Mutah University
College of Graduate studies

Automatic Arabic Text Categorization Using Efficient Classification Techniques

By
Mouhammd Mahmoud Al-awadi

Supervisor:
DR.Mustfa Hammad

A thesis submitted to the College of Graduate Studies in partial fulfillment of the requirements for the Master's degree in Computer Science to the Department of Information Technology, University of Mutah.

Mutah University, 2015

الآراء الواردة في الرسالة الجامعية لا تُعبر
بالضرورة عن وجهة نظر جامعة مؤتة

بسم الله الرحمن الرحيم

MUTAH UNIVERSITY
College of Graduate Studies

جامعة مؤتة
كلية الدراسات العليا

نموذج رقم (١٤)

قرار إجازة رسالة جامعية

تقرر إجازة الرسالة المقدمة من الطالب محمد محمود العوضي الموسومة بـ:

Automatic Arabic text Categorization Using Effcient Classification Techniques

استكمالاً لمتطلبات الحصول على درجة الماجستير في **علم الحاسوب**.

القسم: تكنولوجيا المعلومات.

	التوقيع	التاريخ	
د. مصطفى محمد حماد		٢٠١٥/١٢/٩	مشرفاً ورئيسا
د. احمد بشير الحسنات		٢٠١٥/١٢/٩	عضواً
د. عوني منصور الحموري		٢٠١٥/١٢/٩	عضواً
د. مضر موسى المعاني		٢٠١٥/١٢/٩	عضواً

عميد الدراسات العليا

د. محمد المحاسنة

MUTAH-KARAK-JORDAN
Postal Code: 61710
TEL :03/2372380-99
Ext. 5328-5330
FAX:03/ 2375694
e-mail: dgs@mutah.edu.jo sedgs@mutah.edu.jo
http://www.mutah.edu.jo/gradest/derasat.htm

مؤتة – الكرك – الاردن
الرمز البريدي: ٦١٧١٠
تلفون: ٩٩-٠٣/٢٣٧٢٣٨٠
فرعي 5328-5330
فاكس 375694 ٠٣/٢
البريد الالكتروني
الصفحة الالكترونية

ACKNOWLEDGEMENTS

First, my very special thanks go to my supervisor, Doctor. **Mustafa Hammad** from Mutah University. Working with Dr. Mustafa was a very stimulating and fun experience. In fact he always offers support and feedback positively and gently, without a whimper. Without his support, my the thesis will not be like this.

I wish to express my special thanks to my Doctor. **Ahmad Hassanat**is credited for the starting point of scientific research, he taught me how to be a good researcher, first research paper published by his supervision and encouragement. Dr. Ahmad provides me with support, advices and precious comments over two years,who contributed several ideas on similarity measures. Sharing his knowledge and expertise in this area.

Big Thanks are also due to my colleagues and friends who were willing to assist me during the corpus annotation and coding phases. Without you this thesis would not be as valuable as is it now.

Finally, I would like to particularly mention **Fatimah Alalem,** then Mohammad Al Jamal, Khalil Msedeen, MohanadAzzedaneen, AlaaMazaidah, Ayman, Fras, ashraf and akramshuibat. Much appreciated are likewise because of the researchers from everywhere throughout the world, who liberally made their papers and datasets accessible.Last, however not least, I am extremely thankful to my family, exceptionally my Mum, who made everything Possible to offer me all assistance to achieve my dream.

Mouhammd Al-awadi

Table of Contents

List of Tables

List of Figures

List of Equations

ABBRIVIATIONS

ANN	Artificial Neural Network
ARS	Arabic Light Stemmer
ATC-DS	Arabic Text Categorization Data Set
CCA	Corpus of Contemporary Arabic
CHI	Chi-Square
DF	Document Frequency
DT	Decision Tree
EASC	Essex Arabic Summaries Corpus
ED	Euclidean Distance
EW	Exceptional Words
FN	False Negative
FP	False Positive
HD	Hassanat's Distance
Idf	Inverse Document Frequency
IE	Information Extraction
IR	Information Retrieval
KNN	K–Nearest-Neighbor
LEX	Lexicon
LSA	Latent Semantic Analysis
MD	Manhattan Distance
MI	Mutual Information
NB	Naïve Bayesian
NER	Named Entity Recognition
NLP	Natural Language Processing
OCR	Optical Character Recognition
SL	Stemming List
SVD	Singular Value Decomposition
TC	Text Categorization
TF_IDF	Term Frequency-Inverse Document Frequency
TFIDF-CF	Term Frequency-Inverse Document Frequency Category Frequency
TN	True Negative
TP	True Positive
VSM	Vector Space Mode
W	Word
WEKA	Waikato Environment for Knowledge Analysis

Abstract

Automatic Arabic Text Categorization Using Efficient Classification Techniques

Mouhammd Al-awadi
Mutah University, 2015

Arabic language is a complex language that needs special treatment. However, most previous studies were using statistical methods in Arabic texts classification, and these methods neglect meaning of the terms. Firstly we built an identical Arabic database, so that they are freely available for research purposes in the Arabic language, then designed a framework for preprocessing Arabic text, which consists of multiple steps and modeling techniques, such as stop word removal and a stemmer to improve the results of Arabic texts categorization.

This thesis focuses on the semantics technique, and proposes a hybrid stemmer for Arabic languages. Varies techniques are used to implement the Arabic text classifications, and to verify our hybrid stemmer. These techniques include Latent semantic analysis (LSA) and five machine learning approaches. LSA used to reduce dimensionality in order to improve the accuracy of categorization systems.

The experiment results showed the effectiveness of our Arabic stemmer in terms of classification accuracy and speed. The best performance was achieved by combining Singular Value Decomposition (SVD) with cosine similarity measure and Manhattan distance.Finally, we Compared experimentally; Hassanat's distance with Euclidean's distance, Manhattan distance and cosine distance, to choose the best way to calculate the similarity between vectors with five text representation .

الملخص

التصنيف التلقائي للنصوص العربية باستخدام تقنيات التصنيف ذات الكفاءة

محمد محمود العوضي

جامعة مؤتة، 2015

اللغة العربية هي لغة معقدة وتحتاج الى معاملة خاصة. لهذا، فإن معظم الدراسات السابقة استخدمت الأساليب الإحصائية في تصنيف النصوص العربية، وهذه الأساليب تهمل معنى المصطلح. أولا، قمنا ببناء قاعدة بيانات تماثلية للغة العربية، بحيث تكون متاحة مجان الأغراض البحث العلمي في اللغة العربية، ثم صممنا برنامج المعالجة النص العربي، والذي يتكون من مجموعة من الخطوات وتقنيات النمذجة، مثل إزالة كلمة التوقف وتقليم الجذور لتحسين نتائج تصنيف النصوص العربية.

وتركز هذه الأطروحة على تقنية دلالات الكلمات، وتقترح طريقة مهجنة لتقليم كلمات اللغة العربية. وتستخدم العديد من تقنيات لأغراض تصنيف النص العربي، وعرض نتائج الطريقة المهجنة. وتشمل هذه التقنيات على التحليل الدلالي الكامن (LSA) بالإضافة إلى خمسة من مناهج تعلم الآلة. حيث يستخدم LSA للحد من التشتت الذي من شانه تحسين دقة أنظمة التصنيف.

وأظهرت نتائج التجربة فعالية طريقة تقليم الجذور المقترحة من حيث دقة التصنيف والسرعة. وقد تحقق أفضل أداء من خلال استخدام طريقة تحليل القيمة المنفردة (SVD) مع مقياس جيب التمام ومسافة مانهاتن. وأخيرا، فإننا أجرينا مقارنة تجريبية على مسافة الحسنات،و المسافة الإقليدية،و مسافة مانهاتن و مسافة جيب التمام، لاختيار أفضل طريقة لحساب التشابه بين ناقلات مع خمسة من طرق تمثيل النص.

Chapter 1
Introduction

With the increasing growth of Arabic contents on the internet, the need to organize these contents in a fast and automatic way is also increasing. This environment needs automatic and efficient categorization methods for texts, to be automatically classified into categories and groups. Categorization systems can significantly contribute to monitor huge quantity of Arabic texts that exist in many forms, such as in news, journal articles and web pages.

Generally, people need an easy-to-use approach and retrieval engines to classify files and texts according to their interests. It is hard and difficult task to manually classify documents. The availability of automatic tools that provide such services for Arabic documents are restricted. However, there are many researches focused on some methods, which have more effective morphological analysis and study their effect on Arabic text categorization.(Saad, 2010),(Al-Nashashibi, 2014),(Arabic text preprocessing for the natural language processing applications, 2007)Here, we need more researches to check and improve the accuracy of these methods, moreover, to improve their effect on text categorization (TC) Performance.

Arabic language categorization needs further study and attention, because it is different from other languages, such as English and European languages which have been already developed. There are some problems on the way of Arabic language is structured, such as mismatching obstacle. This issue means that using different words in different texts can give the same meaning as a result. Arabic words have different meanings and usages according to their relations to the context. This gap has been intensively studied over the time and many methods have been examined to solve such problems(Saad, 2010).

Text categorization is one of the artificial intelligence applications(Information Systems Design and Intelligent Applications, 2015), that allows collection a large number of unstructured text to be grouped into classes based on specific criterion, such as topics, author of the text, text language etc. Text Categorization (TC) also called "text classification or document categorization. The interest of Arabic text classification is increased recently.

Text Categorization is overlapping in areas of Machine Learning (ML) and Information Retrieval (IR), and related to many important areas of scientific research, such as data mining, texts retrieval and natural language processing. (Figure 1.1) illustrates the intersection between those fields.

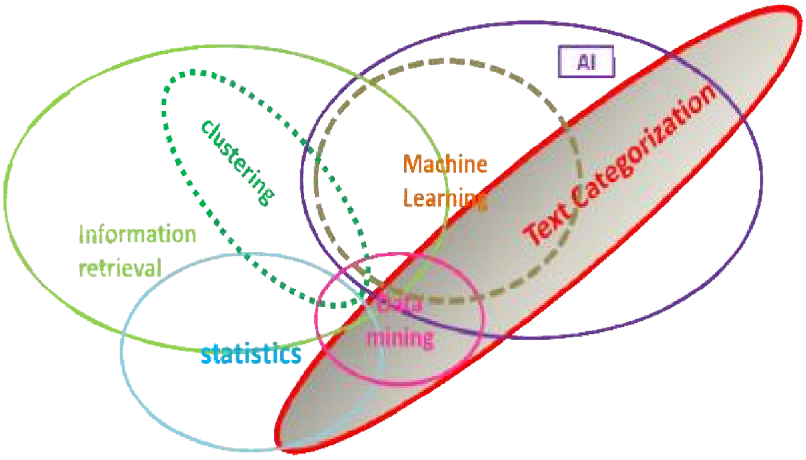


Figure 1.1:
The relationship between Machine Learning, Data Mining, Clustering and Categorization

As shown in (Figure 1.1) the area of categorizing texts intersects with the following fields: Data Mining, Artificial Intelligence and Information Retrieval and Statistics that is separated from clustering.

1.1 Definition of Text Categorization (TC)

TC is the process of identifying a specific text category depending on a set of attributes within that text (Arabic Text Classification Algorithm using TFIDF and Chi Square Measurements, 2014). TC which based on machine learning is composed of training on labeled texts and experimenting on new unlabeled texts. The following diagram (Figure 2) shows the main Idea of Text Categorization.

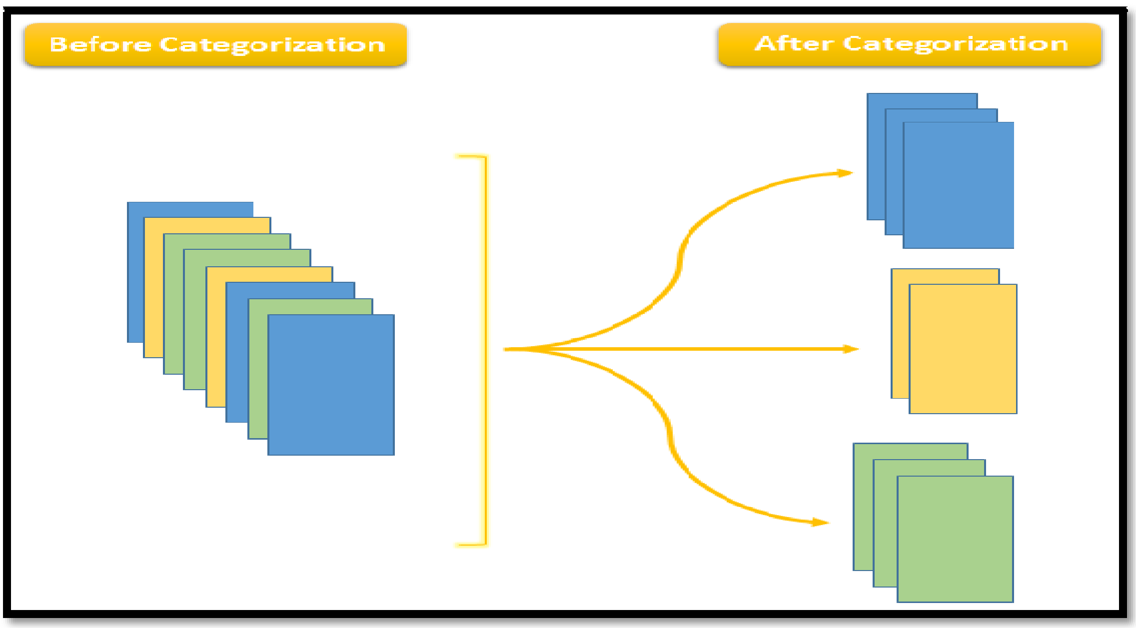


Figure 1.2:
The main Idea of Text Categorization

(Figure 1.2) shows a collection of unsorted documents, which are organized and sorted by content into categories. This is the general idea of texts categorization.

1.2 The usage of Text Categorization

Because of the development of TC, many tools have been appeared. Nowadays, TC is widely used in many applications, such as:

Spam filtering: It is possible to find this term under many of terminologies, such as junk email and unsolicited bulk email. The process is aimed to find the messages that containing unrelated, immaterial and unwanted information to numerous recipient who has not demanded it(Filtering spam e-mail from mixed arabic and english messages: a comparison of machine learning techniques, 2009),(Servers to spam: drop dead. Computers & Security, 1997),(Arabic Spam filtering using Bayesian Model, 2013).These messages are usually sent for the purpose of advertising, promotion, and spreading backdoors or malicious programs. According to recent Statistics (Support vector machines for spam categorization, 1999), 40% of emails are spam, which about 12.4 billion email per day and that cost internet users about $255 million per year.

E-mail filtering: The process of categorization the emails according to their topics or areas (Arabic Alert-Email Detection Using Rule Based Filter, 2011).Every email that we receive containing key words that reflect the topic of it, classifying those emails helps the users to organize and retrieve it when needed(Tracking recurring contexts using ensemble classifiers: An application to email filtering, 2010).

Sentiment Analysis: it is possible to find this term through opinion mining or sentiment mining. Sentiment analysis is the process of identifying trends or decisions of any person relied on his writings. Meanwhile, it follows the mood of the audience through their comments on the product or a specific place that is usually classified into positive, negative and neutral. There is some effort to apply sentiment analysis on Arabic texts, such as (Sentiment analysis for dialectical Arabic, 2015),(Building Large Arabic Multi-domain Resources for Sentiment Analysis, 2015),(Sentiment Analysis For Modern Standard Arabic And Colloquial, 2015).

Routing of Emails: it is the process of choosing the entity in which the message will be sent depending on the topic (mailbox, general contact, and specific contact ….etc.)For example, in major companies such as universities that contains a large number of staff. The administrator needs to send emails to the employees by the nature of their work. Here you can meet this requirement by using the email routing.(Miloslavsky, 2008).

Authorship Attribution: it is the problem of determining the real author's work that is unknown or disputed (Author unknown: On the trail of anonymous, 2000). Paper was presented by Ahmad alfalahi.et al (Identify

author of Arabic poem , 2014)discussed one of the most important Arab studies that used categorization texts techniques for identification of the poet purposes.

Identification of Language: it means the process of distinction between languages and categorization of any text to its language (Arabic, English, Spain,..etc.) (Dunning, 1994).Some studies presented approaches to distinct Arabic language (Approach to determine the language automatically (without supervision) in the city of multilabel texts, 2013),(Arabic Dialect Identification using a Parallel Multidialectal Corpus, 2015).

News monitoring: it is the process of news websites categorization of according to their topics (Efficient monitoring algorithm for fast news alerts, 2007). For example news monitoring can be used to category diseases reports. Moreover Businessmen are interested in news about the economy and stock prices.

Papers categorization: it is also called scientific articles indexing depending on their abstract. The idea of classification of research papers according to their topic or area such as medical, chemical, engineering etc.

Movie Categorization: this is the process that determines whether a movie is preferred, non-preferred, or neutral based on reviews (Movie review categorization based on a multiple classifier, 2007). In addition determining a movie's type, such as action, dramas, or thriller.(Movie classification using visual effect features, 2007),(Automatic categorization of movie trailers using data mining techniques: A review, 2015) .

Classifying websites: Many Arab studies presented Spam Detection algorithms for Arabic Web Pages (Spam detection methods for Arabic web pages, 2012),(Content-based analysis to detect Arabic web spam, 2012),(An Empirical Analysis of Arabic WebPages Classification using Fuzzy Operators, 2009). Also, there are a number of Muslim researchers who used Arabic text categorization techniques in Quran studies. Additionally, it is used to classify Prophet Mohammad's sayings (Peace be upon him)(A Topical Classification of Hadith Arabic Text, 2014). Paper was presented by FauzyHaraq.et al (Feature Extraction of Arabic text classification texts of Prophet Mohammed sayings, 2015)discussed one of the most important Arab studies for the purposes of Extracting features for the categorization of the Hadith texts. In addition to Categorization of Quranic Text (verses)(A Topical Classification of Hadith Arabic Text, 2014),(Searching Concepts and Keywords in the Holy Quran, 2011),(The Usage of Formal Methods in Quran Search System, 2011).

1.3 Arabic Text Categorization Importance

In 1974, Arabic language was the sixth official language of the United Nations. It is also the official language of 27 countries.

Furthermore, it is used by about 480 million people. Moreover it's classified as the fourth most widely used language in the world (2015).
Arabic language is one of the most widespread languages. It is the fourth language proliferation around the world, after Mandarin, English and Hindi. It is used by more than 422 million native speaker (2015). (Figure 3) shows the number of speakers for each language in descending order.

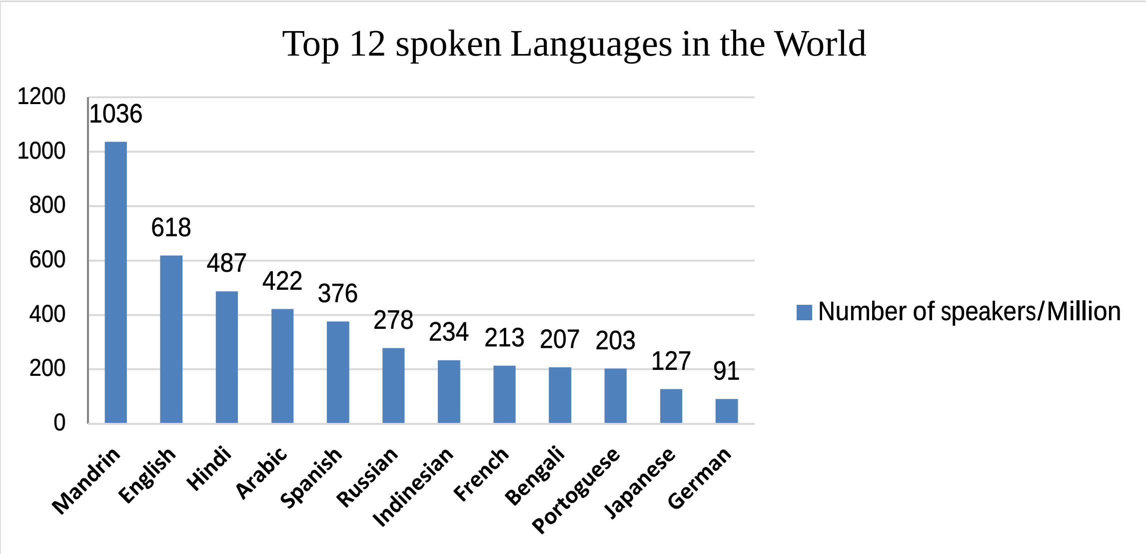


Figure 1.3:

Top 12 Spoken Languages in the World Corresponding to Number of Users.

A study was done on 2014(25/6/2015).Estimates the percentage of websites that contains Arabic contents to be approximately 0.8% of all the contents on the internet. Moreover, Arabic language is not only used by Arabic people, but also by Muslims. This is because all original details for the Muslim religion are written in Arabic language.
The importance of Arabic texts categorization arises for many reasons such as, the large number of the world's population whose mother language is Arabic, The great spread of the on-line documents and social networks, resulting in a lot of Arabic texts on the Internet and the need for treatment and retrieval of such texts. Moreover, Arabic language is a complex language, depends mainly on the context. Another reason is that daily and significant growth of the number of Arabic texts circulating on the Internet. Finally, there is a continuing increase of the number of users of the Arabic language.The complexity of the Arabic language's structure is a reason of lack of automated approaches for Arabic texts classification. The work on this field is started recently.

1.4 Arabic language characteristics and complexity

Arabic Language contains twenty-eight letters with no lower or upper cases. These letters are: (ا ب ت ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع غ ف

ق ك ل م ن ه و ي). Arabic letters are written from right to left as connected letters for one word. In addition to the basic alphabet, Arabic language contains **Hamza** (ء) and **Harakat** (َFatha, ِKasra , ُDammah), which are added to the letters to add a dimension of sound for words, this is because Arabic language relies heavily on Phonics. Arabic language faced a number of difficulties caused by its nature such as:

1. Arabic language lyrics meanings are depending on context, where single-word can carry more than one meaning according to their presence in the text. This issue poses difficulties in classifying texts. For example, the word (حب) has two different meanings (حُب means Love and حَب means beans), another example (عَلَم means flag, and عِلْم means science).
2. There are problems with Arabic language and its privacy, such as forms multiplicity, Harakat (Fatha, Kasra, Damma,...) and historical conflict between mandarin and dialect.
3. Text Categorization (TC) depends mainly on the roots of words that may led to problems in Arabic language. This is because the presences of one root for matching many different words. For example, the root (collect يجمع)shared by a number of words, such as a (group مجموعة), (community مجتمع), (University جامعة) ,(Mosque جامع), (meeting اجتماع) moreover, each of these words has a different meaning.

1.5 Thesis Motivation

The spread of Arabic texts in many computers and on the Internet showed that there is a great need to process those texts and classify them. Furthermore, number of on-line tools and software relating to Arabic language is still limited. In this thesis, a classification approach is proposed for Arabic texts. The proposed approach composed of a novel preprocessing technique, stemmer, and machine learning processes to categorize Arabic texts according to their subjects. The subjects of processed documents may have any type, such as religious, political, medical, economic, etc. the study have taken care of all the factors to get the best results with respect of accuracy and efficiency.

1.6 Research Problems

Most users of computers and information technology think that the relationship between the computer and the language is friendly and simple. Computer helped in the study of human languages and introduced simplified, processed, converted analysis from one form to another. However, experts and specialists cope with it in a different way. Many complexities and ambiguities can be faced. Therefore, the great challenges and risks will be noticed especially phonics languages such as Arabic language.

There are a number of problems encountered during our study including:

1. So far, text categorization has not applied to Arabic texts as in English texts. As a result, there is a lack for tools and frameworks to handle files and documents when dealing with Arabic language.
2. Arabic language has a high sensitivity to context property compared to English language. This means that the characters formed based on the previous characters. This sensitivity in matching Arabic language's components relationships, such as matching the verb with the subject makes Arabic language processing mechanically more complicated and difficult than other languages. This leads to difficulties in finding a mechanism to process Arabic grammars automatically.
3. Lack of enough available databases for Arabic texts to apply the study and test the results. Therefore, this compels us to collect our database manually.
4. There is no standard way to apply the preprocessing steps for Arabic texts in TC. This means that each researcher uses different steps according to his convictions and viewpoints to reach best results. This leads to conflicting judgments between researchers about steps that lead to optimal text classifier.
5. The property absence (Harakat) to the Arabic words can be used either fully or partially. So, this leads to complex confusion situations. The real cause for that is the multiple possibilities for the reading of the word. For example, the Arabic word (وَجَدَ) can be read to signify the meaning of the word (found), or it can be in the form of (وَجَدَّ) which means (tried). The absence of the (Harakat) represents a fundamental obstacle to the Arabic words processing.

1.7Research Objectives

The main goal for this work is to find an efficient and automated methodology to classify Arabic texts according to their subjects. At the end of the thesis, we aim to achieve the following:

1. Build an identical Arabic database, so that they are freely available for research purposes in the Arabic language. This would increase the chances of the detailed studies of Arabic texts.
2. Provide an analytical and comparative study of most well-known preprocessing steps, as well as modified steps, to classify Arabic texts, and compare them to reach the optimal steps.
3. Build a framework for preprocessing Arabic text, which consists of multiple steps and modeling techniques, such as stop word removal and a stemmer to improve the results of Arabic texts categorization.
4. Search for an appropriate way to compute the similarity between two vectors through a series of tests.

5. Perform a study on using some similarity functions between two vectors.

1.8 Research Contributions

This study aims to provide an automatic approach for Arabic text categorization. The proposed approach achieves high accuracy and it is light-weighted with respect of time and memory. Moreover, this study provides a number of achievements in the field of text categorization, which include the following:

1. Identical Arabic text dataset, which we have generated manually. This dataset is balance amount of Arabic texts classified within folders to categories according to text topics.
2. Arabic stemmer for Arabic texts analysis and stemming. This stemmer is compared with other well-known previous stemmers.
3. This work also presents a new technique for Arabic text categorization using Latent semantic analysis (LSA). It used to reduce dimensionality in order to improve the accuracy of categorization systems.
4. What distinguishes this study is that, we presented a comprehensive survey of the steps that precede the Arabic texts categorization, as well as categorization steps, to obtain high accuracy results for categorization of Arabic texts.
5. Compare experimentally; Hassanat's distance with Euclidean's distance, Manhattan distance and cosine distance, to choose the best way to calculate the similarity between vectors.
6. Study the impact of the nature of the used data set in the categorization accuracy, by implementing eleven data sets on six different types of categorization and displaying results.

1.9 Problem Statement

For Arabic language, there is a lack of accurate and efficient classifiers. There are no standard rules to deal with different issues in Arabic language components and structures. The number of studies and researches that contributed to the categorization of Arabic texts is limited. This is due to the nature of the Arabic language, where it relies on phonics. As researchers, we would be interested in a system that can categorize a set of different Arabic texts by providing an accurate stemmer and classifier. Classifying documents automatically saves users' time and effort.

1.10 Research Questions (RQ)

To achieve our objectives from our thesis, we think of several questions. These questions are about the data preprocessing, text

categorization and how to check efficiency ratings. Some of these questions are:

RQ1: We must provide identical database of categorized texts, to apply the experiments. How we will collect this data and what is the best way to number the assembled data?

RQ2: The results of categorization algorithms will be affected by the preprocessing steps. For categorization, stemming is an important step that processes the data before the categorization. How our stemmer will affect the results and the accuracy of the used categorization algorithm? Are all categorization algorithms would be affected by the same percentage, if we use the same stemmer?

RQ3: What are the standard stop words in Arabic language? Can we modify them? Can we ignore them without affecting the classifier's currency?

RQ4: What are the available stemmers for Arabic language? Can we build a modified stemmer based on them?

RQ5: Can we use Vector Space Model (VSM) to represent our data? If yes, what is the best representation? Can the probabilistic models (N-gram) contribute in the representation step to raise the accuracy?

RQ6: The SVD is a method used to get rid of duplicates in your data for the purposes of information retrieval. This technique can be used in the Arabic texts categorization? If possible how can we get utilize from it?

RQ7: In machine learning, there are many used classifiers. What is the best one that can be used in Arabic text classification process?

1.11 Performance Measures

The performance of a text categorization system is usually measured in terms of its effectiveness (Modern Information Retrieval-The Concepts and Technology behind Search, 2011), i.e. its ability to produce correct categorization. Some researchers also compare the performance in terms of efficiency (for example, execution time). Since text collections are typically large, the efficiency of a system is very important in real-life settings. Recall, precision, F-measure and accuracy is the mean four measures used to evaluate efficiency of any classifier. Firstly, we will explain the abbreviations used in the evaluation equations

TP: (True positive): Number of documents that is should be retrieved as related to a category and retrieved.

TN: (True Negative) Number of documents that is shouldn't retrieved as related to a category and not retrieved.

FP: (False Positive): Number of documents that are not related to a category but retrieved.

FN: (False Negative) Number of documents that is not retrieved as related to a category but should be.

Accuracy and Error Rates

Accuracy is the percentage of training texts that assigned to their correct category. Error is the percentage of training texts that assigned to wrong category. Note that these standards become unreal if text in a particular category. Accuracy and Error Rates calculated as follows:

Accuracy Rate: $=\frac{|TN|+|TP|}{|TN|+|TP|+|FN|+|FP|}$…………………………...………(1)

Error Accuracy Rate = $\frac{|FN|+|FP|}{|TN|+|TP|+|FN|+|FP|}$……………………...………(2)

Note that, Error Accuracy Rate = 1 – Accuracy Rate.

Precision and recall

Precision and recall are used to assess the accuracy of information retrieval algorithms, in order to modify the results. Precision and recall in texts categorization are differ slightly from information retrieval, here they are used for measuring the quality of text classifier, and may reduce some problems of accuracy. Precision is the fraction of all texts in a particular category by the classifier that really belongs to that category. Recall is the fraction of all texts that really belongs to a category in and was correctly assigned to same category by the classifier.

Precision $=\frac{|TP|}{|TP|+|FP|}$…………………………………………………(3)

Recall$=\frac{|TP|}{|TP|+|FN|}$…………………………………………….………(4)

F1 Measure: it is used to calculate the harmonic mean of precision and recall. F measure combines the values of precision and recalls into one measure F1 computed as the following formula:

F1-Measure= $\frac{|2TP|}{|2TP|+|FN|+|FP|}+\frac{2*Precision*Recall}{Precision+Recall}$…………….………(5)

We will use the above statistical measurements to judge the efficiency of categorization algorithms that will be used to assess and compare the effectiveness of them.

1.12Thesis Structure

The remainder of this work is organized as follows: in Chapter two we describe the previous works on feature extraction, stemming techniques, categorization algorithms, evaluation, and applications of text categorization. In addition, an overview of previous work on Arabic text

categorization was discussed. Chapter three will introduce the concept of document representation and the most famous classifiers used in Arabic text categorization. In Chapter four, we discuss the preprocessing steps needed to get the data ready for categorization. Then, our proposed method in preprocessing and categorization was introduced in chapter five. The next chapter illustrates the used data set. Chapter seven introduces experimental results and evaluates them with detailed analysis. Finally chapter eight summarizes our study and introduces some recommendations and plans for future works.

Chapter Two:
The Background

2.1Machine Learning

In this chapter, basic fundamentals about classification techniques and documents' representation models are presented.

In ancient times, when large libraries were designed such as, Alexandria library around 300 BC (Kiritchenko, 2005), there was a need to save and retrieve data. With the rapid increase of Documents, the need for efficientmethods for storage and retrieval also increased, day by day texts retrieval become difficult. Researchers had to think about the numbering of books and documents, which increased the organization and simplified search and retrieval. The beginning of categorization was through giving unique numbers to texts, but this method disabled us to retrieve a book according to its subject. This forced researchers to group books and texts by their topics as categories or classes. These categorizations provide information about name, gender, size, language. Etc. The process of adding text to a specific category is called text categorization or text classification. Classification includes text organization and coding, which makes understanding and retrieval better.

Machine learning (ML) (Kodratoff, 2014)is the branch of computer science that is associated with artificial intelligence applications. The term machine learning is based on design algorithms and techniques that resulting in a machine, which has the ability to predict future expectations. Machine learning divided into three main types; supervised learning, unsupervised learning, and semi supervised learning as shown in (Figure 4):

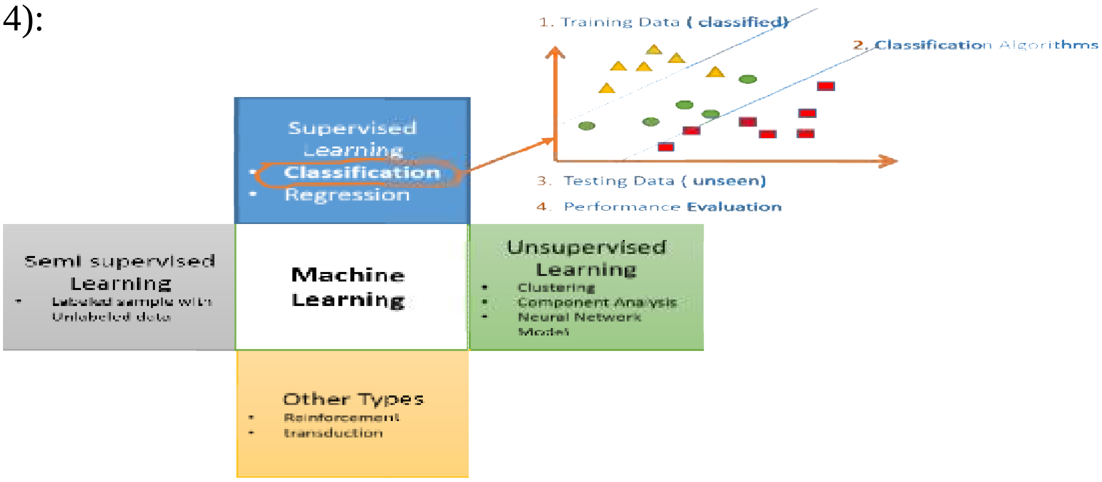


Figure 1.4:
Categorization as Machine Learning Type

As shown in (Figure 4), supervised learning depends on dividing dataset into two parts (Saad, 2010)(Kiritchenko, 2005); training dataset with correct representation, which will be used to train the algorithm. The second part of the dataset is the unclassified data, which is used to examine performance efficiency, depending on applying some measures such as: precision and recall, F-measure. Examples of this type include regression and categorization, which is the core of our study in this thesis.
Another type of ML is the unsupervised learning. In this type, there is no training dataset; all data set is used to get results. Examples of this type are clustering and component analysis. If we group texts into classes without text-labeling, then this operation is called clustering and classes called clusters. The third type of ML is called semi supervised learning where a piece of labeled data set is merged with the rest of no labeled data set. This method reduces the complexity of labeling time. To use supervised algorithms in Arabic TC.

2.2 Representation Models for Documents.

There are two main classes of document representation models, the first class concerned with quantity and the second is quality model. (Figure 5) shows the three models for representing documents.

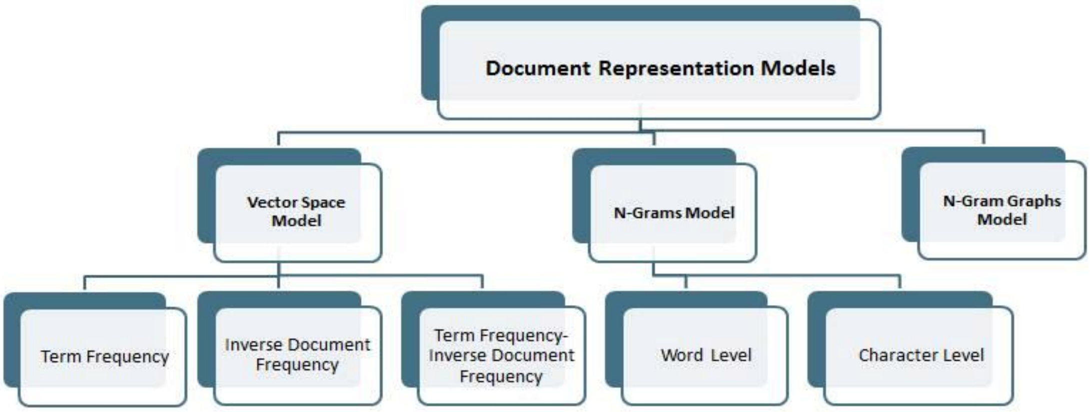


FIGURE 1.5:
Document Representation Models

2.2.1 Vector Space Model (VSM)

VSM is also called vector space model, which introduced by Wong, Salton and Yang for vector representations of texts. This model relies primarily on the number of term frequencies within a document. So, it is sometimes called content-based model or quantity term within text. As shown in (Figure 5), VSM usually can be computed by three models; term frequency (tf), inverse document frequency (idf) and term frequency-inverse document frequency (tf-idf).

The classification of a particular document depends on the frequencies of terms, Ft,d = frequency of term t in document d. This frequency is normalized through dividing the frequency of term by the maximum term frequency in the document. This is called Term Frequency. Term Frequency (TF)(Arabic Text Classification Algorithm using TFIDF and Chi Square Measurements, 2014)model is based on computing the total number of occurrences for a term as follows:

TF t,d = Ft,d/ Max (Ft,d)…………………………………….……..(6)

If (d) denotes the document and (t) denotes the term, then the symbol ($TF_{t,d}$) expresses the number of (t) occurrences in (d).
Inverse Document Frequency (IDF) model (Modern Information Retrieval-The Concepts and Technology behind Search, 2011)is the second weighting schema of VSM that represents the inverse proportion of documents, which contain a term t over the whole set of documents d. df_t is the number of documents which includes term t, whileIDFi is the inverse document frequency of term*i*, It indicates the power of discrimination.
The following formula used to compute idf:

$$IDF\ (d, f)\ =\ log_{10}(\frac{N}{DF(d,t)})……………………………………………………. (7)$$

The third VSM model is the Term Frequency-Inverse Document Frequency (TF-IDF)(Stemming as a feature reduction technique for Arabic Text Categorization, 2011). This model is the logarithm of the proportion of number of documents (N) divided by number of documents containing the given term (n_t) (Al-sarrayrih H, 2009), this method calculates the occurring of the term in a document with no occurring this term in other documents, this means that ratio of discrimination term of the document.
A term occurring frequently in the document but rarely in the rest of the collection is given high weight.
The following two formulas can be used to calculate TF-IDF $_{t,d}$:

$$\text{TF-IDF}_{t,d} = \frac{TF(t,d)LOg(\frac{N}{n}t)}{\sqrt{\sum_{k=1}^{t} TF(t,d)^\wedge 2[Log\ (\frac{N}{n}k)\ ]^\wedge 2}}\ ..…………………………………(8)$$

Or using the equivalent formula;

TF-IDF $_{t,d}$ = *TF (t,d)* Log (N/n t)* ……………..……………………(9)

Where,
TF $_{t,d}$: Number of term t frequency in document d.
N : Number of documents in the corpus.
n_t:Number of documents from the corpus contains t.

TF-IDF assumes that all documents have same length. However, in fact the documents vary in length from one to the other. Therefore, this assumption reduces the accuracy of categorization. In 2012 Lio and Yang proposed TF-IDF model with **category-frequency** by adding a length normalization factor to the TF-IDF formula (An improvement of TFIDF weighting in text categorization , 2012). This factor acts as a category frequency factor (c). The factor c determines term frequency for documents with category C. formula 10 used to compute TFIDF-CF:

$$TFIDF - CF = \log\left(tf_{t,d} + 1\right) * \log\left(\frac{N+1}{n_t}\right) * \left(\frac{n_{ct,d}}{N_{ct}}\right) \ldots\ldots\ldots\ldots\ldots..(10)$$

Where:

$n_{ct,d}$: Number of documents where term I founded in category c.

N_{ct} : Number of documents belongs to category C and term t.

Generally, SVM models have some limitations. For example, they do not take into account the order of words in the text, instead of that they consider only existence and replication of terms in the documents. Moreover, SVM models face the problem of ignoring the context of term or words of similar meaning (Representation Models for Text Classification: a comparative analysis over three Web document types, 2012).

2. 2.2 N-gram Model

Iqbal AbdulBakidefines the N-gram as "a sub-sequence of N-items in any given sequence, where the grams are characters of words"(Classical Arabic Poetry Categorization Using N-Gram Frequency Statistics, 2010). There are two types of N-gram model, character-level N-gram, and word-level N-gram. The first type depends on the sequence of letters where word-level N-gram relies on the sequence of distinct words. Unlike Vector Space Model, N-gram model concerned with term context (Representation Models for Text Classification: a comparative analysis over three Web document types, 2012),(N-Gram-Based Techniques for Arabic Text Document Matching; Case Study: Courses Accreditation, 2012). As a result, the order of a term in the document affects the value of N-gram. The value of N can be choosing in any way that guarantees high efficiency of categorization results. The value determines the number of characters or terms. If N=1, then it is called unigram, if N=2 then, it is called bigram, for N= 3 it is trigram (Classical Arabic Poetry Categorization Using N-Gram Frequency Statistics, 2010), and so on. The following example illustrates the N-gram in the case of word-level for Arabic texts:

The Arabic sentence"الأرض موطني والسماء غايتي" can be represented:

- In Bi-gram (2-gram) as :

  الأرض موطني ، موطني والسماء ، والسماء غايتي ، غايتي.

- Or in Tri-gram (3-gram) as:

  الأرض موطني والسماء ، موطني والسماء غايتي ، والسماء غايتي.

2.2.3: N-gram graphs models: not to be used in our work.

2.3 Categorization Classifiers

All Arabic texts undergo the following steps when classified by machine learning methods:

1. Inputs: Documents divided into two sets, one for training the classifier, which already has been classified, and the unclassified set to test the classifier's accuracy.
2. Specific number of given categories: ctgry1, ctgry2 ... ctgryM
3. Processing algorithm according to the used classifier Output: trained classifier

   The most well-known supervised algorithms that can be used to build a machine-learning based classifier are Decision Tree (DT), Naïve Bayesian (NB), Support Vector Machine (SVM), K Nearest Neighbor (KNN), and Artificial Neural Network (NN) algorithms. The following subsections give a brief description for each one.

2.3.1: Decision Tree (DT)

One of the categorization algorithms that belongs to the learning machine family is decision tree, it is used heavily in text categorization (Automatic Arabic text classification, 2008),(Different Classification Algorithms Based on Arabic Text Classification: Feature Selection Comparative Study, 2015). Decision trees can be used in different text categorization applications such as: spam filtering to determine whether the inbox is spam or legitimate, email routing to send an email to general contact or specific according to its subject, writer genre to determine whether a text written by male or female, etc. many researchers prefers decision tree because it's easy to understand and use, need few simple computations. On the other hand, it's not suitable for many categories or small dataset. The following figure shows a simple decision tree;

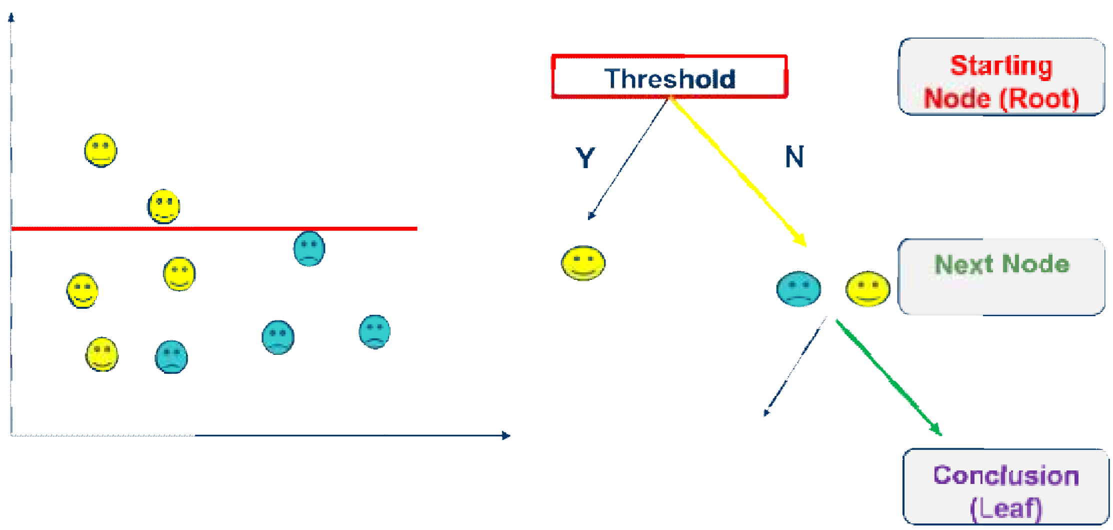


Figure 1.6 :
Decision Tree for Text Categorization

As shown in (Figure 6), to define the best threshold, we must take into account the clarity scale for each node in order to improve the selection of the features. Decision trees classify documents starting at the root of the tree and moving through until reaching the leaf node which indicates the categorization result of a document.

2.3.2: Naïve Bayes (NB)

There are many studies that used NB in categorization (Using Word N-Grams as Features in Arabic Text Classification, 2015),(Automated Arabic Text Categorization Using SVM and NB, 2011),(Different Classification Algorithms Based on Arabic Text Classification: Feature Selection Comparative Study, 2015).When applied to build probabilistic model for each category, depending on the training data set, we calculate the doubling possibilities for all features extracted from the text to be classified so as to give a rapprochement with the text of the anonymous testing. The highest value of the probability that a value of 1 between all categories in our experience are likely to define the class that is the attribution text, some probabilities may yield a result of zero because none of the training data are located in the same range. To avoid zero values, the equation that represents Naïve Bayes is:

$$\text{P(Category/Document)}=\frac{\text{P(Category).P(Document|Category)}}{\text{P(Document)}} \ldots\ldots\ldots\ldots\ldots. (11)$$

P(Category/Document): Is the probability of a category given a document , i.e. the probability that a given document belongs to a given category.
P(Document): The probability of a Document.
P(Category): The probability of a category, to be computed from the number of documents in the category divided by document number in all categories.
P(Document|Category): Is the probability of a given document in a certain category.

2.3.3: Support Vector Machine (SVM)

Support vector machine is a form of machine learning introduced by Vladimir N. Vapnic. SVM first used in TC by Joachims(Arabic Text Classification Using Support Vector Machines, 2009). The basic idea of SVM is to find the best hyper plane separating two categories with the largest possible margin according to the previously classified set of data as shown in (Figure 7). After selecting the appropriate hyper plane, we can use it to divide the data into two categories based on the side of this data.

We can use SVM to cases where the boundary between categories of type nonlinear.

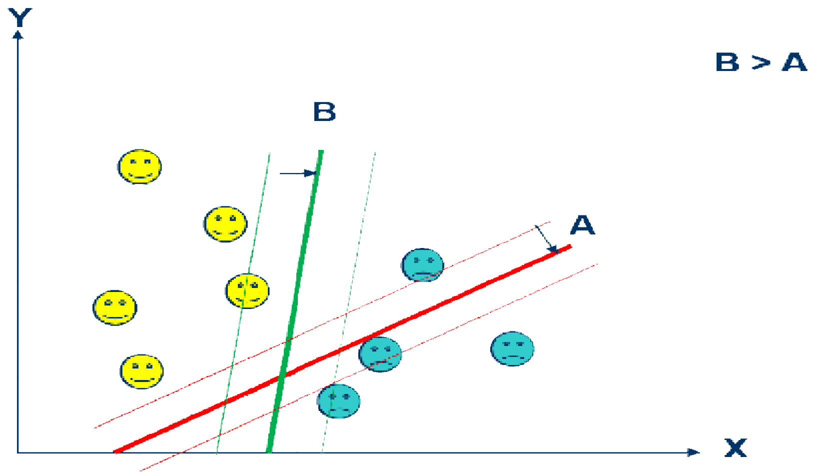


Figure 1.7:
Support Vector Machine Classifier

As shown in (Figure 7), it is clear that the margins of B hyper plane is greater than the margins of A, thus, the best choice is the green hyper plane. SVM has achieved great results in text categorization (Stemming as a feature reduction technique for Arabic Text Categorization, 2011),(Using Word N-Grams as Features in Arabic Text Classification, 2015),(Arabic Text Classification Using Support Vector Machines, 2009),(Arabic Text Categorization Algorithm using Vector Evaluation Method, 2015),(Text Document Classification based-on Least Square Support Vector Machines with Singular Value Decomposition, 2011), so most of the researches for these purposes are based on SVM.

2.3.4: K Nearest Neighbors (KNN)

This is a simple algorithm for text categorization depending on N vectors (Automatic text categorization using neural networks, 1988). Suppose that C is a document to be classified, K will determine the number of nearest neighbors to C. (Figure 8) illustrates an example where K takes different values.

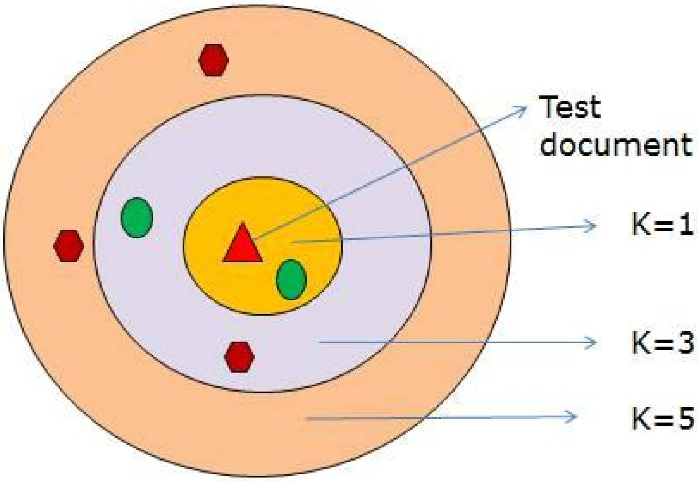


Figure 1.8:
K nearest Neighbors Classifier

In the previous example, when the value of k = 3, then, we will discuss in the nearest 3 neighbors for C, we find that the nearest three neighbors are B, B, A. As B>A Then the vector C belongs to B category. (Using Word N-Grams as Features in Arabic Text Classification, 2015),(Arabic Text Classification Using Support Vector Machines, 2009),(Different Classification Algorithms Based on Arabic Text Classification: Feature Selection Comparative Study, 2015),(Vector Space Models to Classify Arabic Text, 2014) are examples of researches used KNN in Arabic texts categorization.

2.3.5: Artificial Neural Network (ANN)

Neural Network was appeared to simulate human neurons as follows: we can build a network operating on the same principle as neurons, so we can get this smart capable computer, and started the artificial neural network (ANN). It is begun to be used in text categorization in 1995(Arabic NLP tools for ontology construction from Arabic text: An overview, 2015). ANN based on building three layers; input layer, hidden layers and output layer. Input layer consists of vectors after applying feature dimensionality reduction, hidden layers is the activity region in categorization. Output layer represents the categories. Researchers in recent years have tended to use ANN in Arabic texts categorization (Feature Extraction of Arabic text classification texts of Prophet Mohammed sayings, 2015) ,(Stemming as a feature reduction technique for Arabic Text Categorization, 2011) and(Arabic Text Classification Based on Features Reduction Using Artificial Neural Networks, 2013). (Figure 9) shows a simple representation for ANN.

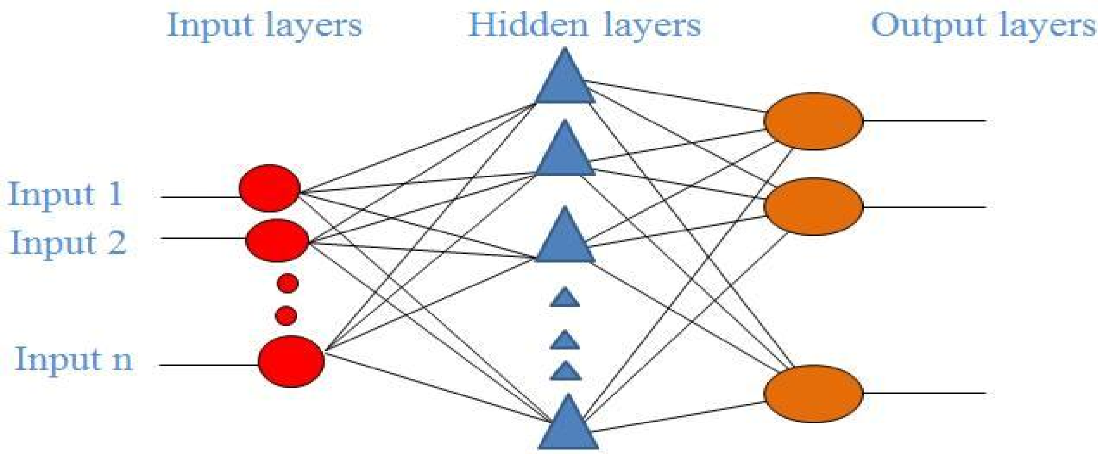


Figure 1.9:
Neural Network Classifier

2.4. Related work

Many studies have been conducted regarding the categorization of texts. Most of these studies were applied on English texts. The characteristics and details of Arabic language (mentioned in chapter one) made a few researchers to be interested in conducting studies on the Arabic texts. This chapter discusses the related work on four aspects of text categorization: Applications of Text Categorization to other domains such as Email Filtering (The impact of preprocessing on text classification, 2014), Author Gender (An extensive study of the Bag-of-Words approach for gender identification of Arabic articles, 2014), etc. Feature Extraction, Categorization Algorithms and Performance Evaluation.

2.4.1 Text Categorization Applications

Although the widespread of the Internet and the existence of too many texts, there is an urgent need to categorize those texts to facilitate the retrieval and utilization of them. Many researchers have employed categorization of texts in various fields such as, Spam Filtering(The impact of preprocessing on text classification, 2014),Sentiment Analysis, Routing of Emails, Authorship Attribution, Identification of Language, Papers and Movie Categorization, Classify web sites, etc.

In (Automatic text categorization and its application to text retrieval, 1999)the researchers designed a way to classify texts intended to retrieve information. Additionally, they found that the categorization texts elevate the efficiency and speed of information retrieval. They also did their experiments on two sets of data on medical topics: HERSH and OHSUMED). Here information can compare the results with and without text categorization.

Liang-Chih and his colleagues (Evaluation on text categorization for mathematics application questions, 2013) used the categorization of texts in mathematics to classifying examinations of various levels of difficulty to

help students to choose questions that fit their level. They found that the decision tree algorithm is the best categorization method (78.53%) compared to other categorization algorithms.

Also, In 2015 Zinner, Thomas, et al. (Text Categorization for Deriving the Application Quality in Enterprises Using Ticketing Systems, 2015)found that Text categorization can be useful in text filtering for identifying relevant support tickets. Because of a very large number of tickets used daily, there is a need for large number of workers to classify tickets which contain problems, researchers thought of an automatic way of categorization for those tickets. They also used 12,000 tickets to apply three text categorization algorithms (NB, KNN and SVM), during the experiments to compare categorization algorithms it turns out that SVM is the best in terms of accuracy. Sentiment Analysis becomes one of the most important applications of categorization texts but this application faces many challenges for the machine to understand human intention.

Emma Haddi and colleagues (The role of text pre-processing in sentiment analysis, 2013)confirmed the importance of text processing by increasing the efficiency of sentiment analysis in terms of precision and recall. Also, they implemented SVM algorithm and compared results before and after processing of texts in three different ways of preprocessing. Another interesting application of text categorization is Author gender authentication.

K. Smairat et. al.(An extensive study of the Bag-of-Words approach for gender identification of Arabic articles, 2014) Introduced a new method for identifying writer gender of an article applying text categorization. They also apply various classifiers on their dataset that contain 500 Arabic articles divided into two classes: male writer and female writer. Their results showed that SVM is the best classifier for this application followed by KNN under specific settings.

Svetlana Kiritchenko(Kiritchenko, 2005)focused on one vital application of hierarchical TC in bioinformatics. This means that he studied three problems: implementing TC in their solution, the first problem indicates biomedical articles indexing, 2nd one is the annotation of gene and the 3rdis the analysis of a given biological information based on clustering. He also introduced a new performance evaluation measure. Islamic Studies were not excluded of the need for categorization algorithms. So, some researchers have used Arabic text categorization to classify Holy Quran texts, and Prophet Mohammed Hadiths by topics as in (Improving Arabic Text Categorization Using Neural Network with SVD, 2010), ,(Stemming as a feature reduction technique for Arabic Text Categorization, 2011). In 2014 Uysal and Gunal(The impact of preprocessing on text classification, 2014) apply text categorization on four types of data to study the impact of

language on categorization , where they used two email sets, one in English and the other in Turkish, to classify emails into Spam or non-spam.

2.4.2 Text Categorization preprocessing

Many studies and researches (Saad, 2010),(The role of text pre-processing in sentiment analysis, 2013)confirmed on the role that was played by preprocessing texts related to the accuracy and speed of the categorization. The nature and details of the Arabic language (Saad, 2010) need ongoing processing. Consequently, many researches focused on studying different ways of texts preprocessing and the impact of these methods to access to optimal categorization. Additionally, most of those researches were done by generating hybrid preprocessing techniques (Stemming as a feature reduction technique for Arabic Text Categorization, 2011)to get high percentage of precision and recall in text categorization. Text Categorization preprocessing involves many steps starting with tokenization and ending with feature extraction. In(The role of text pre-processing in sentiment analysis, 2013)they implemented three preprocessing techniques and compare results. They used feature frequency (FF), feature presence (FP) and Term Frequency Inverse Document Frequency (TF/IDF) with support vector machine classifier. As a result of that, the results showed that using of TF/IDF is the best in terms of performance measures.

Accordingly, Saad, Motaz (Saad, 2010) proved practically that feature extraction, stemming term indexing etc. play a vital role in text categorization in terms of accuracy and speed. Also, he showed that light stemming can reduce the features from 30 to 40%, so this can increase the performance. He also applied his thesis by using the seven Arabic datasets. The results show that SVM and NB algorithms are the best classifiers. From another side, light stemming is also the best in reducing features.

(Support vector machines based Arabic language text classification system: feature selection comparative study, 2008) Shows that using feature selection with CHI square has the best results on SVM classifier. In (Saad, 2010)his thesis, he concluded that reducing features is a vital step to increase the efficiency of the classifier. Also he mainly used light stemmer and term pruning to increase the performance of NB and SVM classifiers. Additionally, he showed that distance-based classifiers are affected by term weighting schemas. (A new and efficient stemming technique for Arabic Text Categorization, 2012)Developed their stemmer according to the three famous Arabic stemmers, and studied its impact on naïve Bayesian classifier. (Arabic text preprocessing for the natural language processing applications, 2007) Used four preprocessing steps (Tokenization, light stemming, morphological analysis and annotation) for NLP. He showed that these preprocessing phases are efficient with

vowelized documents. In relation to that, in (The impact of preprocessing on text classification, 2014) they studied the impact of preprocessing on texts categorization accuracy, where they use four preprocessing steps: stop words elimination, lower case conversion, tokenization, and stemming.

Al-thubaity and his colleague (Using Word N-Grams as Features in Arabic Text Classification, 2015) studied the effects of N-gram on Arabic TC. They used 1,526 documents belong to 6 categories. So, their results showed that the using of unigram is better than 2-gram or 3-gram on NB, SVM and KNN classifiers. Also, the using of SVM classifier yielded in 75.44% of accuracy when applying 1-gram and choosing 200 features. From another side, Dina and Nayer(A study of text preprocessing tools for arabic text categorization, 2009)suggest some tools that will benefit ATC. In this way, they noticed that the use of stemming and morphological tools such as MORPHO3 will affect positively on the categorization process. Meanwhile, the term of weighting is an important step to represent text as vector term. In relation to that, many studies have used the schemas Chi square and TFIDF (Automatic Arabic text classification, 2008),(Arabic Text Classification Algorithm using TFIDF and Chi Square Measurements, 2014),(Arabic Text Classification Using Support Vector Machines, 2009)for term weighting. Aymen Abu-Errub used both methods in the categorization step.

In (Arabic Text Classification Using Support Vector Machines, 2009) the researchers use Chi square and TFIDF for representing text vectors. (Stemming as a feature reduction technique for Arabic Text Categorization, 2011)The impact of using different stemming schemas, root stemmer, light stemmer and dictionary based stemmer were addressed. Consequently their findings show that light stemmer is the best method to enhance categorization efficiency. Also, their experiments were accomplished using Hadiths dataset with SVM and ANN classifiers. This means that this Paper (Automatic Learning of Arabic Text Categorization, 2015) has adopted the principle of similarity measurements, so they used several methods to measure the percentage of similarity; Manhattan, Euclidean Distance and Dice measure, as well as using n-gram through different values for n. After testing, they found that the best results were obtained when n = 3.

2.4.3 Text Categorization

In (Table 1), the researcher summarizes some previous studies findings that concern with Arabic texts categorization. Also, each of these studies used one or more of the famous categorization algorithms such as (support vector machine SVM, K-nearest neighbor KNN, Naïve Bayesian (NB), Naïve Bayes Multinomial (NBM), artificial neutral network ANN, Decision Tree (DT) etc.). In relation to this, all of these studies and

experiments on text categorization need data preprocessing and some efficient steps to get the best results. To assess the results of the text categorization, the researchers adopted texts on a number of metrics including (execution time, precision, recall, F-measure, etc.)

Table 1.1:
A Summary of some Previous Researches in Arabic Text Categorization

#	Paper name	Used Classifier	Preprocessing steps	Accuracy measures	Accuracy percentage	Dataset
1	Arabic Text Categorization Algorithm Using Vector Evaluation Method (Arabic Text Categorization Algorithm using Vector Evaluation Method, 2015)	vector evaluation method words frequency	162 stop word, Normalization, Root stemming, Frequency weight, Max 2 words	Accuracy	93% - 98% depending on category	8112 different words
2	Automated Arabic Text Categorization Using SVM and NB (Automated Arabic Text Categorization Using SVM and NB, 2011)	SVM, and NB	Remove the digits and punctuation marks. Normalization Non Arabic texts were filtered. Arabic function words were removed	Recall Precision F1	SVM P= 0.779 R= 0.778 F1= 0.778 NB P= 0.741 R=0.74 0.74	5121 Arabic documents of different lengths that belongs to 7 Categories
3	Different Categorization Algorithms Based on Arabic Text Categorization : Feature Selection Comparative Study(Different Classification Algorithms Based on Arabic Text Classification: Feature Selection Comparative	(D.T) (KNN) (NB) (NBM)	Feature selection methods with light stemmer, (khoja stemmer) , Feature selection with full word.	Precision recall, F-Measures accuracy time to build model	The worst is KNN. Different results obtained	BBC Arabic dataset 4763 documents belongs to seven categories

	Study, 2015)					
4	استخراج الميزات لتصنيف نصوص الاحاديث الشريفة(Feature Extraction of Arabic text classification texts of Prophet Mohammed sayings, 2015)	SVM Neutral Network	Light stemming Root stemming Dictionary stemming	F1	Neutral Network is butter than svm. In NN dictionary stemming is the best. In svm, light stemming Is the best.	453 document belongs to 14 categories.
5	Arabic Text Categorization Based on Features Reduction Using Artificial Neural Networks (Arabic Text Classification Based on Features Reduction Using Artificial Neural Networks, 2013)	Artificial Neural Network (ANN)	features reduction methods	Precision Recall Accuracy F-measure	Compare results between propagation ANN and ANN Results before and after feature reduction	2800 document belongs to 7 categories
6	Arabic Text Categorization Using Support Vector Machines (Arabic Text Classification Using Support Vector Machines, 2009)	K-NN Rocchio Bayes SVM	Stop words removal Stemming Document indexing Categorization algorithms	Accuracy Time	Rocchio classifier is outperforms k-NN in time and accuracy. SVM outperforms in high dimensional feature spaces	1132 document in 6 categories 22347 unique words from 3 newspaper s
7	Automatic Arabic Text categorization (Automatic Arabic text classification, 2008)	SVM C5.0	Feature selection and term weighting using Chi-Squared	Accuracy	SVM=68.65 C5.0=78.42	17,658 document belongs to 7 categories.
8	Using Word N-Grams as Features in Arabic Text	naive Bayes k-nearest neighbor support	stop words removal numbers and non-Arabic letters removal	Accuracy	Using 1-gram feature resulted in best results. SVM with 1-	1,526 document belongs to 6

	Categorization (Using Word N-Grams as Features in Arabic Text Classification, 2015)	vector machine	Using 1, 2, 3 and 4-gram as features.		gram gives best accuracy (75.44)	categories Saudi Press Agency(SPA) dataset
9	Improving Arabic Text categorization Using Neural Network with SVD(Improving Arabic Text Categorization Using Neural Network with SVD, 2010)	Artificial Neural Network Classic SVD	Remove less than 3 byte words. Some suffixes were removed. stop words, very high and very low frequency words were removed SVD	precision recall F-measure	ANN model using SVD is better than the basic ANN Precision +SVD= 55% Precision without SVD = 53%	453 Documents belongs to 14 categories "Hadiths'
10	Vector Space Models to Classify Arabic Text(Vector Space Models to Classify Arabic Text, 2014)	KNN(text2text comparison) VSM Variations (cosine, dice and jaccard)	Remove digits and punctuation, Normalization Remove non-Arabic texts. Remove stop words.	Recall, Precision F1	Cosine categorize outperformed Dice and Jaccard.	5121 Arabic documents belongs to 7 category Saudi newspapers

As shown in (Table 1), all researchers aim to achieve one goal. The purpose is to increase the effectiveness and speed of Arabic text categorization. So, to achieve this goal, they tried to modify some categorization systems operations. Firstly, the researchers tried to control preprocessing steps of the texts. The whole text is required to delete stop words, digits and non-Arabic words; and to conduct normalization and different stemming operations this subject will be covered in chapter (4). Some claim that light stemming gives the best results (Stemming as a feature reduction technique for Arabic Text Categorization, 2011). Other researchers used root stemming (Arabic Text Categorization Algorithm using Vector Evaluation Method, 2015) or a hybrid of root and light together(Feature Extraction of Arabic text classification texts of Prophet Mohammed sayings, 2015).

On the other hand, a few researchers tried to use the N-gram (Stemming as a feature reduction technique for Arabic Text Categorization, 2011), (Using Word N-Grams as Features in Arabic Text Classification, 2015)as a step to reduce the dimensionality of features. Then, texts will become ready for categorization. Here, each researcher tries a classifier that resulted in best efficiency. Most researchers accepted that using SVM (Stemming as a feature reduction technique for Arabic Text Categorization,

2011), (Using Word N-Grams as Features in Arabic Text Classification, 2015),(Arabic Text Classification Using Support Vector Machines, 2009),(Arabic Text Categorization Algorithm using Vector Evaluation Method, 2015),(Text Document Classification based-on Least Square Support Vector Machines with Singular Value Decomposition, 2011) gives much accurate results than other categorization algorithms. But the others use different categorization methods such as NB (Automated Arabic Text Categorization Using SVM and NB, 2011), KNN (Arabic Text Classification Using Support Vector Machines, 2009), (Different Classification Algorithms Based on Arabic Text Classification: Feature Selection Comparative Study, 2015), (Vector Space Models to Classify Arabic Text., 2014)], ANN (Feature Extraction of Arabic text classification texts of Prophet Mohammed sayings, 2015),(Stemming as a feature reduction technique for Arabic Text Categorization, 2011),

Accordingly, some studies have used more than one categorization algorithm (Automatic Arabic text classification, 2008), (Automated Arabic Text Categorization Using SVM and NB, 2011), and (Different Classification Algorithms Based on Arabic Text Classification: Feature Selection Comparative Study, 2015) to study a particular factor. For example, Alsaleem Saleh(Automated Arabic Text Categorization Using SVM and NB, 2011) implemented NB and SVM classifiers on different Arabic datasets. So he found practically that the SVM yielded better precision over NB by 2.8% and it had to recall and f-measure. Ghazi Raho and his colleagues (Different Classification Algorithms Based on Arabic Text Classification: Feature Selection Comparative Study, 2015) apply four classifiers; DT, KNN, NB and NBM to study the efficiency of each. After applying the four algorithms using their data, they found that a KNN is the worst classifier regarding to accuracy, precision, F1 and recall measures, they use two feature selection methods: chi square and information gain to compare the results of each.

From another point, some Muslim researchers implement text categorization algorithms to classify the Holy Quran and al-Hadith (prophet Mohammad sayings" PBUH"), so for example (Improving Arabic Text Categorization Using Neural Network with SVD, 2010), (Feature Extraction of Arabic text classification texts of Prophet Mohammed sayings, 2015)use 453 Hadiths belong to 14 category in their experiments. Both two papers use ANN algorithm for classifying Hadiths. In relation to this,(Feature Extraction of Arabic text classification texts of Prophet Mohammed sayings, 2015) when you compare ANN results to SVM, you will find that ANN is more efficient under certain settings using F1 measure. In (Improving Arabic Text Categorization Using Neural Network with SVD, 2010)the complete different approach has been proposed by Fouzi, and Al-Qawasmah to improve the performance of Arabic TC.

They use Singular Value Decomposition (SVD) with neural network (NN) algorithm to reduce dimensionality as a method to classify Arabic texts faster and more accurate.in this way, their results show that using SVD with NN increase the precision of categorization from 52% to 55% on average.

In addition to that, Murty et. al.(Text Document Classification based-on Least Square Support Vector Machines with Singular Value Decomposition, 2011) Presented their algorithm (LS-SVM) by improving support vector machine classifier. This means that they split the input matrix into three matrices using SVD decomposer as a way to increase speed and efficiency of categorization.

In 2012 (Text classification by aggregation of SVD eigenvectors. In Advances in Databases and Information Systems, 2012) researchers addressed a method to classify texts according to topics and easiness. So, their findings proved that using SVD and modifying SVD will increase the efficiency of categorization. Here they where they test their method according to two data classes: real data and generated data. During long search in the previous studies, you notice that very few researchers use SVD method in text categorization, especially Arabic texts. For this reason, the current efforts are standing to merge the SVD decomposer with Cosine formula to improve speed and accuracy of Arabic text categorization.

Other experiments have been conducted using SVD in text categorization by Wang and Deng 2009 (A New Text Categorization Method Based on SVD and Cascade Correlation Algorithm, 2009). Choosing this way was motivated by an interest in the relationship between the terms because SVD generates a very large number of the dimensional vectors. Thereby helping to create great relationships between the terms and Dimensional reduction significantly. Their study linked SVD and cascade correlation algorithm to extract features and classify texts efficiently, with average accuracy rate ranged from (90%) in Automobiles category to (92.5%) in Computer category. (Arabic document similarity analysis using n-grams and singular value decomposition, 2015) In 2015 Ashraf S. Hussein proposed a new technique for document similarity using N-gram with SVD. Hemerged N-gram and SVD to build an effective tool to detect the percentage of similarity between Arabic texts. He used Experiments has demonstrated that using SVD with n-gram when n = 1 gave the best result for the examination of the similarity of text through the data set consisting of twenty Arabic texts from Egyptian newspaper. Another study conducted by researchers Cheng and Soon (Neural network for text classification based on singular value decomposition, 2007) about using SVD and neural network in text categorization; they proved same results as searchers of (A New Text Categorization Method Based on SVD and Cascade Correlation Algorithm, 2009).

2.4.4 Performance Evaluation

The performance of text categorization system is usually measured by the terms of its effectiveness (i.e. its ability to produce correct categorization). Additionally, some researchers compared the performance in terms of efficiency like running time. Because text collections are typically large, the efficiency of a system is very important in real-life settings. Also, the consistent comparisons can be tricky relating to the differences in computational resources, algorithm implementation, etc.

Consequently many several evaluation metrics can be assessed by the effectiveness of categorization system. The most common ones are accuracy, error rate, precision, recall, and combined measures of precision and recall. As stated in (Table 1), most researchers use the accuracy as one of the most important measurements that judge the quality of the categorization. Another measurement has been proposed by (Arabic Text Classification Using Support Vector Machines, 2009),(Automated Arabic Text Categorization Using SVM and NB, 2011),(Text classification by aggregation of SVD eigenvectors. In Advances in Databases and Information Systems, 2012)is complexity time. It means that the time required during categorization execution. Researchers here (Arabic Text Classification Using Support Vector Machines, 2009) apply their hypotheses on four different categorization algorithms to compare execution times. The algorithms are: KNN, Rocchio, Bayes, and SVM. In relation to that, those algorithms use different number of terms.

In (Text classification by aggregation of SVD eigenvectors. In Advances in Databases and Information Systems, 2012) the researchers evaluated results in many ways. Firstly, they calculated the time required to build a categorization system to each category individually. Secondly, they calculate the testing time of the classifier which belongs to the interval (Filtering spam e-mail from mixed arabic and english messages: a comparison of machine learning techniques, 2009),(Identify author of Arabic poem , 2014) seconds. Finally, they compute the traditional performance measurements; recall, precision, error rate and accuracy for main categories and sub-categories categorization.

Chapter Three:
The Proposed Arabic Stemmer: Arabic text Pre-processing

3.1. Text Pre-processing

The natural language processing (NLP) is one of the most important topics in the areas of computer science, which is mainly depending on the use of artificial intelligence techniques(Saad, 2010).The main purpose behind the natural language processing is to process the human (natural) languages and understanding it(Arabic text preprocessing for the natural language processing applications, 2007). Human languages are very complicated, for this reason we need an analytical study of the syntactical rules and semantics of the language, as well as, morphological and science of phonology.

NLP is used in many applications(Saad, 2010), (Arabic text preprocessing for the natural language processing applications, 2007) including: machine translation, automatic text, summarization, reference resolution, discourse analysis, machine translation, morphological segmentation, named entity recognition (NER), natural language generation, natural language understanding, optical character recognition (OCR), part-of-speech tagging, parsing, question answering, relationship extraction, sentiment analysis, speech recognition, speech segmentation, topic segmentation and recognition, word segmentation, word sense disambiguation, Information retrieval (IR), Information extraction (IE), speech processing, text correction and document analysis.

All the applications listed above use a set of pre-processing procedures on the original data, as preliminary practice, and applying them to the entire text. Data preprocessing describes all steps of processing that applied on raw data (original) to make it ready (more appropriate) for the next step of processing. What occurs at the preprocessing stage is converting the data format of the original form, which contains many impurities, symbols, letters, pronouns and other high-repetition that do not benefit from their presence to another form that is more convenient for the future processing. The main goal is to get high accuracy and less processing time in a more easily way.

The phase of Arabic text preprocessing is a major challenge. This is due to the characteristics of Arabic language, and it is considered as a crucial stage and plays a main role in the accuracy of any categorization algorithm. The improvement of the preprocessing phase will affect positively the accuracy of text categorization systems.

For pre-processing steps, there are many different algorithms (Text classification for authorship attribution analysis, 2013) and techniques. For example, sample selection (Enhancement Of Arabic Text Classification Using Semantic Relations With Part Of Speech Tagger)(Stemming versus

light stemming as feature selection techniques for arabic text categorization, 2007) is based on choosing a subset that reflects the best representation of the large data set. Transformation (A novel root based Arabic stemmer, 2015) converts original data as single input. Moreover, normalization (Comparative analysis of Arabic stemming algorithms, 2013),(A new and efficient stemming technique for Arabic Text Categorization, 2012) is converting data into a form that is more easily to reach the best accuracy. Another example is noise reduction technique, which is set to reduce the noise and unwanted appendages in the original data. Finally, Features Extraction (Text classification for authorship attribution analysis, 2013)(Stemming versus light stemming as feature selection techniques for arabic text categorization, 2007)is extract the most frequent words in all the documents that are most likely reflect the document or its category. (Figure 10) summarizes preprocessing steps used in our stemmer.

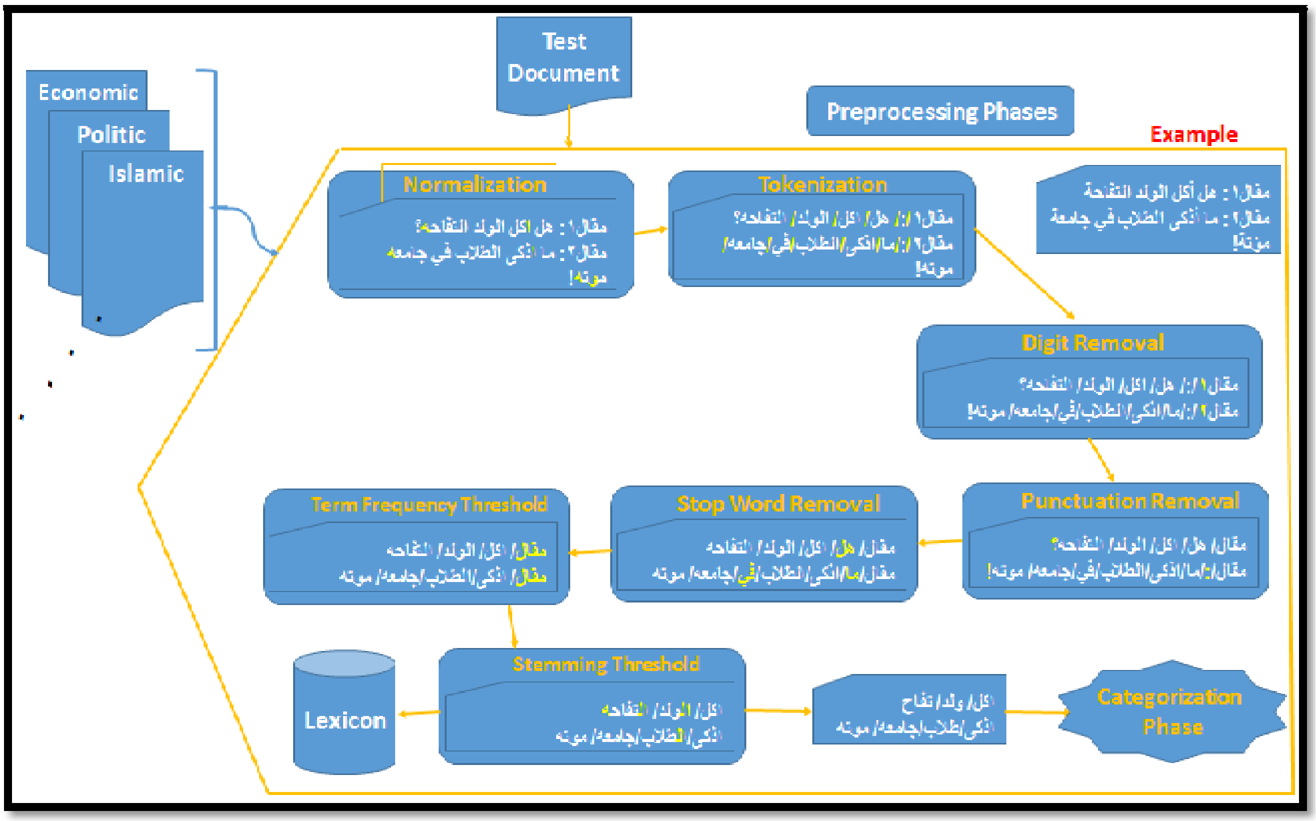


Figure 1.10:
The Main Steps in the Proposed Stemmer (preprocessing steps)

Our approach consists of combining well-known preprocessing algorithms with some modifications to improve the classifier's accuracy. The proposed approach, as shown in (Figure 10), composed of seven steps

for preprocessing the input text. The first step is based on normalization technique (Comparative analysis of Arabic stemming algorithms, 2013)(A novel Arabic lemmatization algorithm, 2008). In this step, the input text is converted to a UTF-16 encoding file. This encoding process is necessary to transform some alphabets to standard forms. Tokenization (Text classification for authorship attribution analysis, 2013)(Comparative evaluation of arabic language morphological analysers and stemmers, 2008) is performed in the second step. This step is set to convert the text files to a series of tokens separated by white spaces. The third step is digits removal, which is implemented by detecting digits' Unicode. The fourth step is punctuation removal, which removes punctuation marks, dashes, special symbols, and any special mark. Arabic Stop words that carry no information, such as prepositions and conjunctions are removed in the fifth step. The sixth step is Term Frequency Thresholding; in this step we will remove all the terms with frequencies that exceed the upper limit of the threshold or the ones that less than the lower limit of the threshold. The last step is Stemming Arabic Words (Text classification for authorship attribution analysis, 2013), which can be defined as the removing of any affixes (prefixes, infixes, suffixes) from the original terms to get new terms called stems (Arabic text preprocessing for the natural language processing applications, 2007)(A new and efficient stemming technique for Arabic Text Categorization, 2012). The following subsections describe each step in details.

3.2. Normalization

Normalization process includes two phases (A novel Arabic lemmatization algorithm, 2008). In the first phase, the input file is converted to a new text file in UTF-16 encoding. This phase can be implemented by using UTFCast express tool ().This tool converts a set of Unicode or ANSI texts to UTF-8 extension. The second phase in the normalization process is standardization, which is set to represent data in a way to reduce the number of terms used by converts certain characters in standard form. In the proposed approach we employed the following normalization steps that are implemented based on (A novel Arabic lemmatization algorithm, 2008) work:

1. Replace the initial letter (“Alif”, " Maddah " , أ ، ا ، آ ، إ) to (bare Alif,ا)
2. Replace the final letter” *tā’ marbūṭah*” ة with “bare haa” ه
3. Replace final letter) “Yaa” ى،ء،ي (with (ى)
4. Replace the letter (“Waaw” " ؤ"،"و ") with ("bare Waaw", (و)
5. Replace the duplicate letter (“double Alif” , أا ، اأ) with ("bareAlif", ا).
6. Replace the long letter (“tatweel”السيــــــــــــاسة) with (السياسة)

(Table 2) shows some examples on normalization process.

Table 1.2 :
Examples of Text Normalization for Arabic Words

Word	Recoding Rule From	To	Word after normalization
يختبئ	ئ	ى	يختبى
سبأ	أ	ا	سبا
إيمان	إ	ا	ايمان
آمال	آ	ا	امال
بؤساء	ؤ	و	بوساء
سيارة	ة	ه	سياره

3.3. Tokenization

Tokenization process is one of the most important steps that used to study text linguistically. Tokenization (Text classification for authorship attribution analysis, 2013) ,(Sentiment Classification on Arabic corpora: preliminary results of a cross-study., 2012) is the method of separation entire sentences into separated words, called tokens. This step can be implemented by finding words separators, such as a white space.

Strings often consist of white spaces, punctuation marks, a number of subscripts and superscripts that include font-size changes, and special characters. The main purpose of applying this process as defined by Bennett et. al. (Stemming versus light stemming as feature selection techniques for arabic text categorization, 2007) is to identify limits of sentences, and to break strings into a set of single tokens by eliminate punctuation.

We assumed that an Arabic token is a series of Arabic letters and diacritical marks that begins and ends with a letter or diacritical marks without separators. Algorithm 1 illustrates the concept of tokenization process. The algorithm starts with the corpus as input, then split the document into sentences and finally sentences into words (tokens).

Algorithm 1: Tokenization Algorithm

Input: The Corpus
Output: tokens (W1, W2, …..)
==

Begin

N: Number of Documents in the corpus.

M: Number of Sentences in each Document.

Dx: xth document in the corpus.

Sy: yth sentence in Dx.

For X = 1 to N

Separate D into sentences
For Y = 1 to M

Separate Sy into tokens (W1, W2, …..)
End

Usually, tokenization process is implemented at the word level to benefit from the meaning of words. There are three common methods to implement this process: bag-of-words, N-gram, and lexicon Word Net.

The bag-of-words (Combining Bag-of-Words and Bag-of-Concepts representations for Arabic text classification, 2013), (Enhancement Of Arabic Text Classification Using Semantic Relations With Part Of Speech Tagger), (Impact of Stemmer on Arabic Text Retrieval, 2014) method is based on separating input string into bag of single tokens. This method is used to convert texts to vectors in a simple way.

The second method is the N-gram method (A new and efficient stemming technique for Arabic Text Categorization, 2012). In this method, the token is considered as a series of n words in the input text. It may be one, two, three, four or five words (Integrating effective rules to improve arabic text stemming, 2014). When the size of n is equal to one, this is called unigram. When the size of n is equal to two, this is called bigram. When n=3 it is a trigram. Greater sizes called by the size of n, e.g., "6th gram", "4th gram", etc.

For example, in the sentence « أضاع اللاعبون جهود ثمينة », when we use the unigrams « أضاع»، « اللاعبون »، « جهود»، and « ثمينة». In English its equivalent to “Players”, “lost”, “valuable” and “efforts”. The bigrams that we can find are « أضاع اللاعبون »، « اللاعبون جهودا » and « جهودا ثمينة» (after

English translation “Players lost”, “lost valuable”, and “valuable efforts”). We will use n-gram in the proposed approach.

The third method is lexicon Word Net (Text classification for authorship attribution analysis, 2013) ,(Enhancement Of Arabic Text Classification Using Semantic Relations With Part Of Speech Tagger).Arabic Word Net AWN (Enhancement Of Arabic Text Classification Using Semantic Relations With Part Of Speech Tagger)is a database of Arabic-language words that is called the group of synonyms, the main objective of the AWN is for natural language processing. AWN was built according to methods developed for EuroWordNet. According to 2008 statistics, Arabic language has 18,957 expressions 6252 nouns 2260 verb, 606 adjective and 106 adverbs. One of AWN formats is XML format, which consists of Items, Words, Links and Forms tags. The main concept of AWN is represent texts as vectors using sum of frequencies for all synonyms.

3.4. Digit Removal

We assume that digits are irrelevant to the text subjects and can be ignored. It is possible that the original documents contain a mix of letters and digits. These digits are deleted by referring to their Unicode representation.

3.5. Punctuation Removal

It is the process of elimination of punctuation marks, such as Full stop (.), Comma (,), Semicolon (;), Colon (:), Apostrophe (’), Hyphen (-), Dash (–) Brackets (() , [ ]) , Inverted commas(" '), Exclamation mark (!), Question mark (?), and any other special symbols (<,>, :,{, },[,] , ^,&,*,(,) , | etc, from the Arabic text.

What distinguishes the Arabic language from other languages is the presence of Harakat and Tashkeel. There are many types of Harakat such as:

1. Fatha, Damma and Kasra(◌ِ ◌َ ◌ُ)
2. Tanween (◌ٍ ◌ٌ ◌ً)
3. sokoon(◌ْ), indicates silence letter.
4. Shaddah (◌ّ) indicates letter duplicate.

These diacritics play a key role in identifying the meaning of a word. In such instance, we cannot discern the word without diacritic. For example, the word (جَد) with fathaa means grandfather, but with Kasra (جِد) means working serious. Our results are not affected by this problem significantly, because most texts we dealt with are long.

3.6. Arabic Stop words Removal

Stop words are a set of repeated words which do not have information, such as pronouns, prepositions, conjunctions etc.(Building an Arabic Stemmer for Information Retrieval, 2002),(Impact of Stemmer on Arabic Text Retrieval, 2014).Words that does not give information to the classifier is ignored. Excluding of such words affects data processing positively.

Other studies have the same results, such as (Comparative evaluation of arabic language morphological analysers and stemmers, 2008). As in the following table (Table 3), this shows the consuming time to classify the data before and after the stop word removal.

Table 1.3

The Impact of Arabic Stop-Word Removal

Name of Database	Build time before stop word removal	Test Time before stop word removal	Accuracy before stop word removal	Build time after stop word removal	Test Time after stop word removal	Accuracy afterstop word removal
ACT-DS1	5.04	13.46	.83	2.58s	7.6	.87
ACT-DS2	1.9	5.53s	.71	0.9s	3.37s	.72
ACT-DS3	2.7	8.73	.85	2.53	8.12	.88
ACT-DS4	3.86	9.14	.86	1.71	4.39	.88
ACT-DS5	.75	2.71	.37	0.68	2.71	.42
ACT-DS6	5.72	7.65	.91	5.68	6.57	.93
ACT-DS7	1.08	1.86	.94	1.02	1.57	.95
ACT-DS8	1.39	2.64	.90	1.3	2.07	.91
ACT-DS9	40s	55s	.84	35.37	40.03	.85
ACT-DS10	1.69	3.8	.93	1.39	3.87	.95
ACT-DS11	1.68	7.58	.73	2.1s	6.63s	.75

We apply different data sets using Bayesian classifier before and after removing the stop words. As you can see from the table, there was a clear impact on the speed and accuracy of results after the removal of stop words, e.g. in ATC-DS11 dataset the accuracy of the results increased by (2.2%). When applying on ATC-DS1 database, the categorization accuracy affected positively by (3.7%), this indicates that the type of data base also affects the accuracy of the results. ATC-DS9 dataset required 55 seconds to test the model before removing the stop words, while it took 40.03 seconds after removed, earning 14.97 seconds. It is noted that the accuracy of the results dataset ACT-DS5 was clearly affected by eliminating stop words and in others has not changed much like ACT-DS2.This is due to the

number of the stop words in documents, but not monitoring any negative impact to remove the stop words.

In Arabic language there are many words not considered as part of speech (Arabic Spam filtering using Bayesian Model, 2013) such as prepositions (الى، عن، في،...،) demonstratives (هذا، هذه، هذان ،...) adverbs (فوق، تحت،...، ... etc.) Days of week (السبت الاحد الاثنين ..) months of yearفبراير مارس . (...) All these words are considered as Stop words. However, there is no specific list or standard list of excluded stop words. Therefore, that this step is considered essential in most of the studies which dealt with texts.

In this thesis, we collect and build 19464 stop words and characters for the Arabic language. These stop words consist mainly of prepositions , interrogation articles, exception articles, demonstratives, condition articles, relative pronouns, verb nouns, sounds nouns, addition nouns, the five nouns in Arabic, emphasis letters, explanation articles, " kaanawakhwatuha ", " innawakhwatuha", catching up letters, expectation letters, selection letters, simile letters, time adverbs, place adverbs, answer letters, assertion letters, vocation letters, relative and isolated pronouns, "kadwakhwatuha", letter nouns, months names, number ordinal noun currency noun, and day noun.

Several stop words from previous studies are added to our stop words list. For example, (Classification of Arabic Documents, 2012) provides 192 stop words that are being freely available on link (2012) within their project. Additionally, in (Effects of stop words elimination for Arabic information retrieval: a comparative study, 2006) 1377 stop words is proposed. Also, those words were checked by Khoja (S. Khoja, 1999) and Alshehri's(Optimization and effectiveness of n-grams approach for indexing and retrieval in Arabic information retrieval systems, 2002) lists. Meanwhile, they used English sources, such as Standard English lists, the Okapi and SMART lists. From another point of view, the following online website added time stop words. This website in (2011) involves two lists: the first one that was collected in 2011 including 162 and the second that was collected in 2014 including 162 too. From this link (), you can download excel sheet which includes 12201. It is available since 30-09-2012. (Table 4) shows a sample of our stop words list.

Table 1.4
A Sample of Arabic Stop Words

Prepositions							Extra word					
امس	اليوم	غدا	نصف	حقا	ثمن	سدس	خمس	ربع	ثلث	سدس	تسع	به
انا	هم	هذه	كل	خلا	بعدما	قبل	ما	تلك	بين	لدن	هيهات	عن
انت	هي	ايانا	عرض	حاشا	قدام	وراء	نحو	بيننا	سيما	حينما	كي	او
انتم	اولئكم	اياكم	قد	اذ	هناك	اعلى	حيث	لهن	حسب	قط	اين	مثل
انتما	هو	اياي	مقابل	الا	حوار	امام	فوق	بينما	ربما	رب	نعم	اياكما
انتن	اولئك	اياكن	سنوات	اذا	هنا	تحت	حيثما	كيف	حسبما	ايان	اجل	أي
بل	عنها	فهو	اعادة	عدد	عدة	عندما	عند	احد	اثر	باسم	شخصا	اطار
بن	هاتان	قال	الاخيره	ثلاثه	ايضا	ايار	اكثر	من	مجاور	كن	كل	في
حوالي	خلا	الثانيه	ثلاث	اثنان	اثنتان	احدى	مائه	عشر	الان	ثمان	سبع	الف
صفر	خمس	اربع	ليست	مابرح	ما فتئ	منفك	صار	سيما	اللذان	امسى	امست	اصبح
ضد	يلي	التالي	الوقت	المقبل	مساء	الذين	الاولى	دون	كنا	نحو	اوضح	اكد
ضمن	الذاتي	جميع	الوكاله	مايو	لن	قوة	كاكم	ولكن	لدى	زيارة	اي	هذان
غير	بان	كما	ظل	بات	ليس	كأن	ليت	إن	أن	دام	ما برح	زال
كان	اصبح	امسى	غروب	معظم	اغلب	جزء	شروق	طلوع	اول	اخر	ادبار	كم
كثير	قليل	بعض	ستكون	بدلا	لدي	ابو	سوف	زلت	فقط	ثم	لهما	جدا
مازال	ولا	لا زال	ما يزال	وقد	منذ	هن	تلكم	قوة	حيث	يمكن	لكن	وقف
مقابل	لقاء	نفسه	برهه	ليال	ليل	ظهر	عشاء	بكرة	مساء	صباح	ساعه	سنه
مليار	بدايه	نهايه	بدل	عدا	بعد	خلف	كيفما	ذلك	ريثما	حين	هاهنا	لما
نحن	هما	هذا	محاذى	خلال	بلا	دونما	تلكما	التى	اتى	مذ	متى	على
قالت	إضافة	اللاتى	اللواتى	بليون	كذلك	لا	يوم	واحد	له	لها	بجانب	كانت
هولاء	اللتان	اياك	بالقرب	لايزال	حول	إنه	بلى	يكون	أصبحت	لم	السابق	حتى
البارحه	بها	لهم	أضاف	عام	ولم	حيثما	مليون	حاليا	عشرون	أحد		

By eliminating the proposed stop words from the input text, our experiments' results show that the processing time can be decreased up to 50% of the original time. Furthermore, accuracy is increased in all test cases.

3.7. Thresholding of Term Frequency

In this step we will remove all the terms of frequencies that exceeded the upper limit of the threshold, or less than the lower limit of the threshold (Light stemming for Arabic information retrieval, 2007). In other words, any term with few numbers of occurrences in the same category's files will be ignored. On the other hand, any term that recurs frequently in all categories is skipped.

This step helps to improve the categorization accuracy and increase the speed, because terms that have a frequency less than the lower threshold, indicate that rarely existence words in the documents. So, they do not express any category. Also, common terms with high-frequency in almost all the documents and all categories. These do not carry

discriminative information, thus cannot be utilized in the process of text categorization.

Removing as much as possible of terms that do not positively affect the system performance will help to improve the efficiency and reduce the time needed for categorization. There is no theory rule to calculate the threshold, but it is calculated by trial and error (empirically).So, it depends on the used datasets. In this step, we determine the upper and lower threshold, and then delete all the words less than the lower threshold and all the words that exceed the upper threshold. This step helps to improve the categorization accuracy and increase the speed, because terms that have a frequency less than the lower threshold, indicate that rarely existence words in the documents. So, they do not express any category. Also, common terms with high-frequency in almost all the documents and all categories. These do not carry discriminative information, thus cannot be utilized in the process of text categorization.

"N-lower Threshold" means that we used a lower Thresholding where N is the size of the threshold. The size of N maybe 1, 2, 3 …. The idea is to remove the frequency less than the lower threshold where frequency means number of appearance in the corpus. To express the frequency rate, we shall use the following formula:

$$\text{Frequency Rate} = \frac{\#\ \text{FT}}{\#\ \text{TC}} \quad \ldots\ldots (12)$$

Where # FT is the number of frequencies of the term, and # TC is the number of text files in the corpus.

Empirically, through experiments we found that the best size of N threshold can be when the size of N is 1(i.e. 1-lowerThreshold) where it is filtering all the words that appear once in the corpus. Means that in the case, depending on the previous equation that the ratio will be which tiny percentage devolves to zero.

As for the upper threshold, we will delete any repeated term in all the texts file in the corpus, this means that the number of appearances in the corpus (#FT) equals the number of text files in the corpus (#TC) based on the previous equation appears to be 100%.

3.8. Arabic Word Stemming

Many languages around the world, such as the Arabic language, must use some additional words that are called affixes (Building an Arabic Stemmer for Information Retrieval, 2002). Those affixes add new information or clarify the meaning of the sentence. In Arabic language, this include the prepositions (e.g. بـ، ل) the pronouns (ex.هن، ت، هم) the conjunctions (ex.و، ف،)and the article (ex. ال). The number of those

additions is limited; the type of the affix is determined by its position within the word. When the affix appears in the beginning of a word, it called prefix, when it appears in the middle of a word it is called infix. Otherwise, it is called suffix (Building an Arabic Stemmer for Information Retrieval, 2002)(Arabic Light Stemmer (ARS), 2014). There is no doubt that the stemmer step has a real impact on the effectiveness and evident on the categorization results, and this is what has been proven in many studies that deal with Arabic text, such as (Saad, 2010).

For categorization purposes, removing of those affixes will positively affecting text compilation in time consuming and accuracy. Comparing Arabic language with other languages, such as English, Arabic language is affected more than the other languages after removing affixes (Arabic text preprocessing for the natural language processing applications, 2007)(Word Stemming for Arabic Information Retrieval: The Case for Simple Light Stemming, 2012).This is due to the complexity of Arabic language structure. (Figure 11) shows an example of an Arabic word structure:



Figure1.11
Parts of Arabic Word: An example

In step seven, we count the number of word letters after deducting prefixes, if it is less than three then ignore the deduction process, and delete only suffixes, then calculate the number of characters again. If L is smaller than three working delete prefixes only examine if the number is less than three characters consider the original word as stemmer.

Table 5
The Used Prefixes

	1- letter	2- letter	3- letter	4- letter
Noun	ك	اب اف ال اك كم فم بم لم لل ول وب	بال فال كال وال ولل الت است للت نست فست	وبال وكال المت واست وللت
Verb	ي م ت ل ن ف س	سي ست سن مت وتيت لي لت لن فت في فن فل ول ون ين نت سا ات ان	وسي وسن وست ولا ولي ولت ولن فسي فال فسن	وساف وسيت واست وتست وفسي اتست انست ستستسنست
Article	ب ل ف ا و			

Table 6
The Used Suffixes

	1 letter	2 letter	3 letter	4 letter
Noun	ك ه ة ي ا	ان ين ون ات ها هم هن كم كن ناية يه تك تي	هما كما يات تان	اتهم اتهن اتها تهما
Verb	ك ه ي ت ا	وا تم تن يا وه ان ون ين ها هم هن كم كن نا ته ني	هما كما نكم	ونني ونهم ناها تماه ناهم ناكم
Article	ه ك	ها هن كم ما كن	هما كما	

The stemmers are a pruning procedure of the original textual word by getting rid of appendages and potential affixes, additions and extract the basic form (core word). Many studies such as [Abu Salem et al. 1999; Xu et al. 2002; Moukdad 2006] said that; Arabic language is different from other languages like English, in terms of the removal of affixes not necessarily reflect the same meaning of the word before and after the applying process of stemming. When use the original form of the word, this caused a challenge to researchers who are trying to build a search dictionary.

This is because the number of index terms will be very huge, which reflected negatively the accuracy and consumes a lot of time and storage space. Therefore, using a stemmer reduces number of index terms. Where it expresses many words derived from this term. Also, will reduce the storage memory used and the time it takes to do the job.

There are many criteria and measures that we can use to assess the quality and efficiency of the stemmer algorithm. And measure the impact of these algorithms on the accuracy of texts categorization. (Will be explained later in details)

3.8.1 Arabic Stemmers Reviews

There are two types of morphological analysis techniques used for text preprocessing (Stemming versus light stemming as feature selection techniques for arabic text categorization, 2007),(Saad, 2010)light stemmer (Stemming versus light stemming as feature selection techniques for arabic text categorization, 2007),(Building an Arabic Stemmer for Information Retrieval, 2002), just remove affixes from the term and merges the terms into the closest category. Heavy stemmer, also called sometimes stemmer (Stemming versus light stemming as feature selection techniques for arabic text categorization, 2007),(Word Stemming for Arabic Information Retrieval: The Case for Simple Light Stemming, 2012), when use a variety of root of word and try assign into the actual root. There are a set of general assumptions for any light stemmer: The term basically consists of two

parts: affix and stem. Sometimes the term contains prefixes and suffixes together, other terms may have only one affix or no affixes. The stem has at least 3 letters; this means that any word less than four letters will be left as it is. Length of affix can be (0, 1, 2 or 3) letter.

Heavy stemming algorithm reduces the words to only their stems (Word Stemming for Arabic Information Retrieval: The Case for Simple Light Stemming, 2012). For example, stemming would result in طبخ، the stem for “to cook” from الطباخ “the chef /one who cooks” .Light stemming, however, does not reduce the word just to its stem, but instead removes some common affixes(A novel Arabic lemmatization algorithm, 2008). For instance, light stemming would yield طباخ from الطباخ، here removing the ال for the definite article. Another example is the reduction of الموظفون، “the Employees” to موظف “Employee,” which involves the use of light stemming to remove the definite article prefix and the pluralizing suffix. It is clear that light stemming is better than heavy (Word Stemming for Arabic Information Retrieval: The Case for Simple Light Stemming, 2012),(Stemming versus light stemming as feature selection techniques for arabic text categorization, 2007) stemmer approach in terms of enhancing feature reduction while simultaneously maintaining word meanings (Stemming versus light stemming as feature selection techniques for arabic text categorization, 2007). For this reason, we have used the light stemming facility in our thesis.

There are many problems occur when the removed letters are part of the full word, such as (تسمين) the stem is a wrong word (تسم) after striping the letters (ين) as a suffix. Another example (كالمنا ، بالغ،) the stem is (منا،لغ) after striping the letters (كال ،بال) as prefixes. This means that the light stemmer may change the meaning of the word. In this thesis, we have proposed a solution to this problem by collecting feedback and reviewing the results to determine the special cases that have appeared in the result and excluded when applying the stemmer.

In general, we will summarize the benefits of stemmer (On Arabic search: improving the retrieval effectiveness via a light stemming approach, 2002)(Sentiment Classification on Arabic corpora: preliminary results of a cross-study., 2012) as follows: the stemmer usually expresses a large set of words. Therefore, this would reduce the size of storage space. In addition, it is used to compress documents. It can also be used in spell checking, by searching for the root of the words rather than the original words. This minimizes number of words that make up the dictionary. One example is the text searching using different search engines in the Internet, where search for the root works to match between these roots with the words that resemble it.

There are four main approaches for building Stemmers (Arabic Light Stemmer (ARS), 2014) ,(Comparative analysis of Arabic stemming

algorithms, 2013),(A new and efficient stemming technique for Arabic Text Categorization, 2012); morphological approach, stem-based approach, statistical-based approach (N-gram) and manually constructed dictionaries approach. In the morphological approach, the stemmer extracts the lexical root of Arabic based on morphological analysis (Stemming versus light stemming as feature selection techniques for arabic text categorization, 2007) ,(Comparative evaluation of arabic language morphological analysers and stemmers, 2008). Many of algorithms use this type of stemmers, such as (A novel root based Arabic stemmer, 2015), which is the most famous of these algorithms, called Khoja algorithm, and the most commonly used (Automated Arabic Text Classification with P-Stemmer, Machine Learning, and a Tailored News Article Taxonomy, 2015). It eliminates both the longest prefix and suffix. Then, in order to take out the root, the algorithm links the word that remains to verbal and nominal patterns by using a handful of linguistic data files, to include lists of punctuation marks, stop words, diacritics, and definite articles.

There are three steps in the Khoja algorithm (A novel root based Arabic stemmer, 2015). First, in order to normalize the words, Khoja Arabic Stemmer removes the diacritics, replaces آ أ إ with ا، replaces ة with ه، replaces ى with ي، and once again removes diacritics. The second step is the stemming of the prefixes by removing (ال و ال، بال، كال، فال، لل،)Finally, the third step is the stemming of the suffixes, which removes (ها ان، ات، ون، ين، ية، ه، ي،).

There are a few weak points of using the Khoja Stemmer (Enhancement Of Arabic Text Classification Using Semantic Relations With Part Of Speech Tagger),(A new and efficient stemming technique for Arabic Text Categorization, 2012). One of these weak points is the fact that, in order to ensure new words are stemmed correctly, maintenance of the root dictionary is necessary. A second weak point is that when Khoja Stemmer encounters a vowel letter, it replaces it with letterو which sometimes results to a root that is unrelated to the original word. A third challenge is that, as a result of its specific order of affixes, sometimes not all affixes are removed (Automated Arabic Text Classification with P-Stemmer, Machine Learning, and a Tailored News Article Taxonomy, 2015). Also, the conversion from raw text to vector representation is a challenging task to determine polarity.

The Stem-based approach (A new and efficient stemming technique for Arabic Text Categorization, 2012)depends on the removal of prefixes and suffixes do not depend on the real root of the term. This Stemmer been used by many of the studies (A new and efficient stemming technique for Arabic Text Categorization, 2012),(Arabic text preprocessing for the natural language processing applications, 2007),(A novel root based Arabic stemmer, 2015), (Comparative analysis of Arabic stemming algorithms,

2013). It is worth mentioning that so far there is no specific list of suffixes and prefixes. All algorithms in this field provide a simple series of steps to get rid of a few prefixes and suffixes.

As we have noted that the root-based approaches affect the meaning of the word, on the contrary of this model which removes additions that do not carry any meaning. This method aims to strengthen the categorization accuracy while maintaining the sense of the word (Saad, 2010).

Larkey's light (Light stemming for Arabic information retrieval, 2007) is considered as the most famous stemmer used, relied on stem-based approach. This study proved that the use of this type of stemmer significantly affects the accuracy of data categorization.

That's been somewhat conflicting with the results given by Xu et al. (A novel Arabic lemmatization algorithm, 2008). Where results presented little influence after using this stemmer, knowing that both studies used the same corpus.

In the statistical stemmer approach (N- gram) (A new and efficient stemming technique for Arabic Text Categorization, 2012)(Impact of Stemmer on Arabic Text Retrieval, 2014)(Comparative evaluation of arabic language morphological analysers and stemmers, 2008), related words are assembled using a variety of algorithms that measure the similarity between the letters that make up words. N-gram (Arabic stemming without a root dictionary, 2005),(Al-sarrayrih H, 2009). Considered one of the most famous statistical methods used in this field. Any word consists of a series of letters (N), this is to split words with high match in N - gram. At first glance, such a way may seem unhelpful when applied to the Arabic language, because of the nature of morphological and syntactical analysis. However n-gram method have been used for Arabic languages as mention Mayfield et al. in this study They got good results and significantly improve with 6-gram (Light stemming for Arabic information retrieval, 2007). The fourth approach is manually constructed dictionary approach (Comparative analysis of Arabic stemming algorithms, 2013),(Light stemming for Arabic information retrieval, 2007),(Comparing words, stems, and roots as index terms in an Arabic information retrieval system, 1994), which mainly based on indexing each word to extract roots and stems in order to build the dictionary. Alkharashi(Comparing words, stems, and roots as index terms in an Arabic information retrieval system, 1994)created a small root-stem dictionary with 526 roots and 725 stem.

We will compare between groups of stemmers, each one of these types use a certain set of prefixes and suffixes. According to the results obtained Kanan and Fox (Automated Arabic Text Classification with P-Stemmer, Machine Learning, and a Tailored News Article Taxonomy, 2015) p-stemmer achieved the best results over the five variations of Larkey’s stemmers (Improving stemming for Arabic information retrieval: light

stemming and co-occurrence analysis, 2002) (light1, light2, light3, light8, light10). These results are presented based on the use of three different classifiers: Naïve Bayes (NB), Support Vector Machines (SVM) and Random Forest (RF). There are also different types of stemmers, such as Alstem, Umass and Modified Umass(CLIR Experiments at Maryland for TREC-2002: Evidence combination for Arabic-English retrieval, 2003)where Alstem obtained the best results over Umass and Modified Umass. In 2013 (Comparative analysis of Arabic stemming algorithms, 2013)Otair Showed that the stemmers achieved the best results will pass from five variations of Larkey's stemmer (Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis, 2002)(light1, light2, light3, light8, light10) In addition to Umass and Modified Umass. Work in (On Arabic search: improving the retrieval effectiveness via a light stemming approach, 2002) proposed a novel light stemming technique, results showed that the light stemming technique has achieved great superiority over root-based technique. Also in 2012 Mustafa (Word Stemming for Arabic Information Retrieval: The Case for Simple Light Stemming, 2012)presented simple light stemming algorithm, results have shown that there is a specific set of suffixes and prefixes affect the efficiency rating of Arabic texts.

Taghva and his colleagues (Arabic stemming without a root dictionary, 2005) presented ISRI stemming algorithm using roots and Arabic patterns, is closed to Khoja stemmer but without a root dictionary (A novel root based Arabic stemmer, 2015). The algorithm based on the length of words, pattern, removal of diacritics, prefixes and suffixes. The ISRI achieved good results compared with the Khoja and Light stemmer. (Table 7 and Table 8) show the suffixes and prefixes used in each stemmer.

Table 7

The Prefixes Used in Some Stemmers

Prefix	Light 1	Light 2	Light 3	Light 8	Light 10	P-Stemmer	Alstem	Umass	Modified Umass	Aljilayl	Mustafa	Taghva
ال	√	√	√	√	√	√	√	√	√	√	√	√
ا											√	√
ب											√	√
اب											√	
بت							√					
بال	√	√	√	√	√	√	√	√	√		√	√
بم							√					
ت											√	√
س											√	√
سي										√	√	
ست							√			√	√	
سن											√	
ف											√	√
فال	√	√	√	√	√	√	√	√	√		√	
ك											√	
فم							√					
في							√				√	
كال	√	√	√	√	√	√		√	√	√	√	√
كم							√					
ل					√	√					√	√
لت							√					
لل					√	√	√					√
لي							√					
لا							√					
لم							√					
لن											√	
مت							√					
ن											√	√
نت							√					
و		√	√	√	√	√		√	√		√	√
واست							√					
وت							√					
وم							√					
وا											√	
وي							√				√	
وال	√	√	√	√	√	√	√	√	√	√		√
ول					√	√	√				√	
ولل							√		√		√	√
ي							√				√	√

Name of stemmer: Light 1, Light 2, Light 3, Light 8, Light 10, P-Stemmer, Alstem, Umass, Modified Umass, Aljilayl, Mustafa, Taghva

Table 8
The Suffixes Used in Some Stemmers

Suffix	Name of stemmer											
	Light 1	**Light 2**	**Light 3**	**Light 8**	**Light 10**	**P-Stemmer**	**Alstem**	**Umass**	**Modified Umass**	**Aljlayl**	**Mustafa**	**Taghva**
ا							√				√	√
ات				√	√		√	√	√	√		
ان				√	√		√	√	√			
تم							√					
ان										√	√	
تمل												√
تا											√	
تك							√					
ت											√	√
تين												√
تن												√
ته											√	
تها											√	
تي							√					√
ك											√	√
كم							√			√	√	√
كمل												√
كن												√
ما												√
ن											√	√
نا							√				√	√
ني												√
ه			√	√	√		√	√	√		√	√
ها				√	√		√	√	√		√	√
هم							√			√	√	√
هما											√	
همل												√
هن							√				√	√
وا							√				√	
ون				√	√		√	√	√	√	√	√
وه							√				√	
ي				√	√		√	√	√		√	√
وها											√	
يا												√
ية				√	√		√	√	√			
ين				√	√		√	√	√	√		√
يه				√	√		√	√	√			

3.8.2 The proposed Arabic Stemmer

Our stemmer includes all suffixes and prefixes that are shown in (Table 7) and (Table 8) to increase the accuracy of the TC. **The proposed Stemmer algorithm** is divided into four parts. The first part, which is shown in (Algorithm 2.A), is set to find the exceptions that contains words that not belong to any rule. (Table 9) shows these exceptional words. The second part, which is shown in Algorithm 2.B, deals with weights or patterns of words. Algorithm2.C presents a solution for some words depending on number of letters, and the last part is shown in Algorithm2.D, which is set to remove affixes.

Algorithm2.A: Exceptions

Input: Texts to be stemmed.
Output: Text after stemming (Stemming list)
==============================
Begin
Read W
If W belongs EW then
Add W to SL
Else
GO TO Algorithm2.b
Go to step 1

According Algorithm 2.A. firstly, Read the text Word by Word (W) from the document. Second For each word (W) in the text If W belongs to the table EW then adds W to the Stemming list (SL). Algorithm 2.A deals with words that violate the rules of Arabic language, in the algorithm we will compile the anomalous words that do not belong to the language rules in (Table 9) named Exceptional Words (EW). Then we will input the text to the algorithm and examine every word, if the word from within the anomalous words, then the stemming form is as the same original word. For example, in the sentence:
"ألوان علم لبنان جميلة" (The colors of the flag of Lebanon are beautiful)
Will be dealt with according to the algorithm, as follows:
Is the word (ألوان colors) within the EW-----> yes, Add the original word to the stemming list.

Is the word (علم flag) in the table EW --------- No, then stemming rules will be applied, and the stemming word will be added to the stemming list. Is the Word (لبنان Lebanon) in the table EW ---> yes, Add the word to the stemming list.

Table 9

The Exception Words

		suffix				prefix		
الله	كهرباء	كنيسه	رنين	كمبيوتر	ليل	بكتيريا	مليونا	كبير
كائن	كاسر	كآبه	حنين	كبد	لوعه	علاج	اوروبا	كتف
كثر	كثير	كشف	جنين	كعك	مرجان	لسان	الانترنت	كافر
كفر	كافه	كتاب	لبن	كاتب	لمع	قاهره	اميركا	كره
كوريا	كروي	كرسي	ليمون	كرم	لكم	لحن	سوريا	كهل
كسوف	ايمن	كلمه	لون	كامل	التصق	لثام	عشرين	كتله
شفاه	منبه	مياه	لغه	فكاهه	لحم	لئيم	انترانت	جبهه
لفظ	لطخ	لحس	ماعون	لعاب	لصوص	محمد	بخمسين	لعب
لظى	لفيف	لازب	اسبوع	لطم	لص	لالى	بريطانيا	لولو
لبيب	لبوه	حاله	ايام	معلومات	لحظه	لوح	تلفزيون	لسان
بيانات	سعوديه	سنوات	بغداد	باريس	اليابان	تسعينيات	كرك	يوما
فرنسا	مليون	المانيا	لبنان	رضوان	مجافاه	اجنبيه	كيمياء	كلب
نزاهه	نكهه	فاكهه	كلام	سعاد	رواه	معاداه	مراعاه	اللهم
لون	دوامه	لمح	زكاه	لوعه	مجازاه	معاناه	نجاه	لكمه
لهب	لهجه	ضجه	موالاه	لقح	اجوبه	مكافاه	موالاه	لقيط
موته	بابا	ماما	اب	ابنيه	اتسم	موسيقى	منطقه	بستان
اتصال	اتضاح	اتفاق	توجه	اتوماتيكي	اجاب	اباح	ابيض	احمر
ازرق	اخضر	اسود	اصفر	اجتماع	اجراء	اجنده	احذيه	استياء
ايران	حذيفه	طفيله	معاويه	كونغرس	قانون	كرك	قانون	جامعه

Algorithm2.B: Weights (Patterns)

```
1. Count =0
2. L = number of letters If L = 2 then
3. for i=1 to5
   Stem1 = " و " + 1st letter + 2nd letter
   Stem2= " ا " + 1st letter + 2nd letter
   Stem3 =  1st letter + " ي " + 2nd letter
   Stem4 =  1st letter + " و " + 2nd letter
   Stem5 =  1st letter + " ا " + 2nd letter
If stemibelongs LEX then

   Add stemito SL
   Else
   Add wordto SL
4. If L = 3 then
   Add wordto SL
5. If W starts with "ال" and ends " ،ك",
   Then
   Add wordtoSLwithreplace ئي with Hamza "ء".
6  If W match with Wei or count =1 then
   GO TO Algorithm2.D
   Else
   GO TO algorithm2.C.
```

According Algorithm 2.B, firstly, calculate Length (L), where Length is number of letters in each word. If L = 2 add "و" or ا""in the beginning or add "ي","و" or ا"" in the end. Compare these words that were generated after adding with a lexicon that has been outfitted for this purpose. If any one of these generating words existed in the lexicon returns it otherwise returned the original word (before adding).

We have built a semantic lexicon contains 10,987 words, the words were collected manually through the 11 data set which will be described later in chapter five in description data set. In addition to this data sets relied on two dictionaries available Online werec"Almaany3"(2015)and "the free dictionary4" (2015).

Return to the previous algorithm, if the word consists of three letters keeps it the same. If it ends with"ئي" replaced by Hamza "ء" For example كيميائي(Chemical) become كيمياء(Chemistry) . If it starts with the definite article "ال" and ends "ك ", then "ك" will not considered as suffix. For example (سلوك,بنوك) the word remain the same.

Finally, we check entered words with a list of the most famous Arabic words pattern (Building an Effective Rule-Based Light Stemmer for Arabic Language to Improve Search Effectiveness, 2012) listed in the

following table. Most of these words are taken from Khoja and Marwan stemmers. We created a table using a set of words as weights (Table 10). In this algorithm we got rid of one of the biggest problems. The problem is that some suffixes delete letters from the original word, such as a(كائن) object) instead of becoming (ائن) as in root stemmers, here (فاعل) pattern keeps the word as it. Also, the words كاسر، باسل follow the same rule.
If the word was on the weight of one of the Patterns in Table 10, we go to the Algorithm2.D otherwise, we go to the Algorithm2.C

Table 10

Patterns (Weights) Used in Arabic Language

	Suffix		Prefix		Affix	
يستفعل	فعلان	فعال	تفاعل	انفعل	افتعال	مفعيل
Yastafel	Faalan	Feal	Tafaul	Enfaal	Eftial	Mafeel
يفتعل	فعلاء	مفاعل	مفعل	استفعال	افعلال	فعلل
Yaftael	Falaa	Mafael	Mafal	Istifaal	Ifelal	Falal
تفتعل	فعيل	انفعال	افعال	استفعل	تفعل	فواعيل
Taftael	Faeel	Infeal	Afaal	Istafal	Tafal	fawaeel
فعائل	مفعال	تفعيل	افتعل	افعل	فاعل	تفاعيل
Fa'ael	Mefaal	Tafeel	Iftal	Afal	Fael	Tafaeel
مفاعيل	فواعل	مفتعل	مستفعل	مفعول	فعول	
Mafaeel	Fawael	Muftal	Mustafel	Mafoul	Faoul	
افاعل	فعلى	منفعل	متفاعل	فعاله	متفعل	
Afael	Feli	Munfail	Mutafel	Faaluh	Mutafel	
فعلله	افعلل	افعوعل	مفاعله	متفعلل	افتعلل	
Fallah	Afalal	Ifaual	Mufael	Muftall	Iftaell	
فعلالي	مفعله	فعوله	فاعله	تفعلل	فعالل	
Falali	Mafalh	Faouluh	Faeluh	Tafalal	Faalel	

Algorithm2.C: Affixes –Based

If L(w)>= 4 then
WordP=word-(prefix)
If WordP belongs LEX then

Add WordPto SL
ElseAddwordto SL
Else
WordS=word-(Suffix)
If WordS belongs LEX then

Add WordSto SL
Else
Add wordto SL
If L(WordP) or L(WordS)<3 then
Count ++
GO TO Algorithm2.A

In this thesis we have collected a large number of affixes which are trailing characters in the Arabic language, and not out of the root and remove them, but it was provided that the resulting Word matching a lexicon (LEX) words of Arabic language which we have compiled for this purpose. If it doesn't match, we return the stem to its original state to ensure no distortion; this was logical and efficient solution that's exactly what proved in results. It is noted that some terms were not affected much by stemming process due to the nature of the documents composed to this terms, which would consist of few prefixes and suffixes.

In algorithm 2.C, if Length of Word greater than or equal four then apply prefix removal according table prefix (see Table 5) then checking whether word is exists in lexicon (LEX) or Not. If word after prefix exists in lexicon then add it to the stemming list. Otherwise apply suffix removal according to suffix table (see Table 6). Then checking whether word is exists in lexicon (LEX) or Not. If word after prefix exists in lexicon then add it to the stemming list. Otherwise add original word to stemming list. If Length of Word became less than three then ignore the removing process and add the original word to the stemming list. For example, الوان، سنون.

Algorithm2.D: Letter Based

1. *If L (W)= 4 then*
   If Weight(W) = " افعل " then Weight(W) = (فعل)
   If Weight(W) = " فعول " then Weight(W) = (فعل)
   Add Weight wordto SL
   Else
   Return original word.
2. *If L = 5 then*
   If Weight(W) = " فعالة" then Weight(W) = (فاعل)
   If Weight(W) = " افعال" then Weight(W) = (فعل)
   If Weight(W) = " مفاعل" then Weight(W) = (مفعل)
   If Weight(W) = "افعلة" then Weight(W) = (فعال)
   If Weight(W) = " فواعل" then Weight(W) = (فاعل)
   If Weight(W) = " فعلاء" then Weight(W) = (فاعل)
   If Weight(W) = " فعائل" then Weight(W) = (فعيل)
   If Weight(W) = " فعائل" and 2nd= "و " then Weight(W) = (فائل)
   If W ends with "ايا" replace "ية"
   If W starts with "ا" and ends with "اء" then remove "اء" and add "ي" = 3rd letter

   Add Weight word to SL

   Else return the original word (W)
3. *If L > = 6 then*
   If Weight(W) = " مفاعيل" then Weight(W) = (مفعول)
   If Weight(W) = " تفاعيل" then Weight(W) = (تفعيل)
   If Weight(W) = "فواعيل" then Weight(W) = (فاعول)

If W starts with "ا" and ends with "اء" then remove "اء" and "ا"

Add Weight wordto SL.

Else return the original word (W)
End

If Length of Word (L) equal four then follow weights rule as follows: Weight is " افعل " remove the 1st letter (فعل) and add it to the stemming list, for example "شهر" ، "أشهر" or Weight is " فعول " remove the 3rd letter and add it to the stemming list, for example "، "بروج برج""otherwise return original word.

If Length of Word (L) equal five then and Weight is " فعالة " then convert the word to the weight " فاعل " and add it to the stemming list, for example

"كتابة" , "كاتب" or Weight is " افعال " then convert the word to the weight " فعل " and add it to the stemming list, for example "اشجار" , "أوراق" ، "شجر" , "ورق" or Weight is " مفاعل " then convert the word to the weight " مفعل " and add it to the stemming list, for example "مجالس" , "معامل" ، "مجلس" , "معمل" or Weight is " افعلة " then convert the word to the weight " فعال " and add it to the stemming list, for example "أجهزة" , "أسمدة" ، "جهاز" , "سماد" or Weight is " فواعل " then match the word with the weight " فاعل " and add it to the stemming list, for example "سواحل" , "نوادي" ، "ساحل" , "نادي" or Weight is " فعلاء " then match the word with the weight " فاعل " and add it to the stemming list, for example "غرباء" , "سفراء" ، "غريب" , "سفير" or Weight is " فعائل and 2nd + 3rd + 4th letters are "وائـ" then , remove 2nd letter and add the word to the stemming list for example سوائل becomes سائل. Else convert the word to the weight "فعيل " and add it to the stemming list, for example "حشائش" , "قصائد" ، "حشيش" , "قصيد". Note that the word "جزائر" is in the exceptional table (Algorithms2.A).

If the word ends with "ايا" remove this suffix and add "ية" at the end then add the word to the stemming list, for example, "هدايا" , " ، "هدية وصايا"" , "وصية"If W starts with "ا" and ends with "اء" then remove "اء" and add "ي" before 3rd letter. For example) Writers ادباء) , (اديب writer). Else return the original word (W).

If Length of Word (L) equal six and Weight is " مفاعيل " then convert the word to the weight " مفعول " and add it to the stemming list, for example "مجانين" . "مكاسير" ، "مجنون" , "مكسور" or Weight is " تفاعيل " then convert the word to the weight " تفعيل " and add it to the stemming list, for example "تمارين" , "تقارير" ، "تمرين" , "تقرير" or Weight is " فواعيل " then convert the word to the weight " فاعول " and add it to the stemming list, for example "قواميس" , "قاموس". Or If W starts with "ا" and ends with "اء" then remove "اء" and ""ا add the word to the stemming list as it.

3.8.3 The proposed Stemmers validation

Many researchers designed stemmers for Arabic language; some of them root based, remove prefixes or suffixes, or both. In (Table 11) we show some stemmers and methodology followed, as well as some of the words that stemmers failed to resolve correctly, and the way in which we solve the same words.

Table 11
Solutions of Some Incorrect Results from Other Stemmers

	Name	Methodology	Incorrect examples	Our solution	Match percentage
1	(An accuracy-enhanced light stemmer for arabic text, 2011)	light stemmers	بدائل ،بديلة جزائر جزيرة	بدائل بديل جزائر ، جزائر	90.8% ± 1.59%
2	(S. Khoja, 1999)	Khoja	المنظمات ، ظما البنوك ، نكأ ركبتيه ، ركبتيه بتوجيه ، وجه بستان ،بسأ مرجان رجن محاورات حور سيطرة طرة	المنظمات ، منظم البنوك ، بنك ركبتيه ، ركبة بتوجيه ، توجيه بستان ،بستان مرجان مرجان محاورات محاور سيطرة سيطر	78.44%.
3	(Stemming versus light stemming as feature selection techniques for arabic text categorization, 2007)	new light and heavy Arabic stemmer	أقدام ، قدم للتقارير لتقارير للمستهلك هلك للحواسيب لحواسيب بالمنطقة منطق وصفات صفا تعيينات عين	أقدام ، قدم للتقارير تقرير للمستهلك مستهلك للحواسيب حاسوب بالمنطقة منطقة وصفات وصف تعيينات تعيين	75.03%
4	Ghwanmeh Stemmer (Enhanced Algorithm for Extracting the Root of Arabic Words, 2009)	root-based algorithm based on the Arabic morphological patterns	التراثية راث اجازات جاز مرجان مرج والسيطرة سطر القساءقسأ بالمنطقة نطق	التراثية تراث اجازات جاز مرجان مرجان والسيطرة سيطر القساء قساء بالمنطقة منطقة	67.40%
5	(Arabic Light Stemmer (ARS), 2014)	ARABIC LIGHT STEMMER (ARS)	امتحانات محن موازين موز روابي روب ريان رين مصائر مصر ندرس ندر موسيقى سقى بكتيريا كثر	امتحانات امتحان موازين ميزان روابي رابية ريان ريان مصائر مصير ندرس ندرس موسيقى موسيقى بكتيريا بكتيريا	92%
6	Light Stemmer (Larkey) (Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis, 2002)	Arabic light stemmers	كامل امل مجانيين مجان متفاهمون فاهم فالح ح متمرن مرن بخلاء بخلا بادرودرو	كامل كامل مجانيين مجنون متفاهمون متفاهم فالح فالح متمرن متمرن بخلاء بخلاء بادرو بادر	light stemmers yield much better effectiveness than those that use morphological stemmers
7	(An accuracy-	Root or stem	تمارين مرن	تمارين تمرين	Not available

	enhanced light stemmer for arabic text, 2011)		ماعون معن موازين وزن مصائر صير مضامين ضمن	ماعون ماعون موازين ميزان مصائر مصير مضامين مضمون	
8	(Building a syntactic rules-based stemmer to improve search effectiveness for arabic language, 2014)	rules-based stemmer(Building a syntactic rules-based stemmer to improve search effectiveness for arabic language, 2014)	متلهفون لهف لمحافظون حفظ دوامه دوم فتيان فتي الاطباء طاب العلاج علج المنطقة منطق	متلهفون متلهف لمحافظون محافظ دوامه دومه فتيان فتيان الاطباء طبيب العلاج علاج المنطقة منطقة	Not available
9	(A new and efficient stemming technique for Arabic Text Categorization, 2012)	hybrid method	ركبتيه ركب تستغرق غرق بنقائصنا نقص منظمات نظم	ركبتيه ركبة تستغرق غرق بنقائصنا نقيص منظمات منظم	82.33% (A new and efficient stemming technique for Arabic Text Categorization, 2012)
10	Aljlayl(On Arabic search: improving the retrieval effectiveness via a light stemming approach, 2002)	Light Stemming Approach	بالقاهرة بالقاهرة الاربعين اربع بالشيخ بالشيخ بصفه بصة ومعاركه معارة	القاهرة القاهرة الاربعين اربعين بالشيخ شيخ بصفه صفة ومعاركه معارك	87.4%
11	Mustafa(Word Stemming for Arabic Information Retrieval: The Case for Simple Light Stemming, 2012)	light stemming	بيت بي مساهمات مساهما سمات سما وقت قت مؤلفات مؤلفا	بيت بيت مساهمات مساهم سمات سمه وقت وقت مؤلفات مولف	light stemming Outperformed heavy stemming. About eighteen percent (18%)
12	Taghva(Arabic stemming without a root dictionary, 2005)	A Root Dictionary	القاهرة قهر مناسبة اسب محمد حمد مساهمات سهم	القاهرة قاهره مناسبة مناسب محمد محمد مساهمات مساهم	0.480
13	Our	hybrid method	-----------	رنين كمبيوتر الليل بكتيريا	

Table 11, shows some examples that illustrate some of the solutions provided by our approach. These studies have identified many examples that were considered in a number of studies as a challenge.

For example, in the stemmer presented by EL-BELTAGY and RAFEA (Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis, 2002), they faced problems in processing some words like (بدائل alternatives), (Algeria جزائر) after stemming the word becomes (بديلة، جزيرة)، this result is incorrect. In comparison with our stemmer, we solved the first word"بدائل" it becomes "بديل" By applying Algorithm2.C; if the weight of word is "فعائل" then convert the word to the weight "فعيل." Also for (الجزائرAlgeria), this word is from within the list of exceptions table, and thus it remains the same as the original form.

Khoja's Stemmer suffers from a lot of problems as in the examples referred to (Table 11) and as observed our way provided a solution. The words "المنظمات، ركبتيه، المحاورات" become "منظم، ركبة، محاور" by applying Algorithm 2.d. Also, word "البنوك" According to the algorithm 2.d it becomes بنوك because W starts with the definite article "ال" and ends "ك"، Then "ك" will not considered as suffix and According to the Algorithm2.B Weight is " فعول " remove the 3rd letter. Then the word converts to "بنك." As for words "بستان، مرجان" is from within the list of excluded words. As evident from the table, our approach solves many of the problems that have faced a number of light stemming algorithm such as ARS (Arabic Light Stemmer (ARS), 2014), Larkey(Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis, 2002),Aljlayl(On Arabic search: improving the retrieval effectiveness via a light stemming approach, 2002),Mustafa(Word Stemming for Arabic Information Retrieval: The Case for Simple Light Stemming, 2012).Some of these problems solved by algorithm 2.A such as " قاهرة، موسيقى، بكتيريا، ماعون، منطقة". In addition to some of the problems that have been solved by algorithm 2.B such as "كامل، فالح" and by algorithm 2.C steps 6 and 7. Such as "مصائر اطباء موازين مجانيين،" .In addition, our approach solves many of the problems that have faced a number of root-based algorithm stemming such as Taghva(Arabic stemming without a root dictionary, 2005), Ghwanmeh(Enhanced Algorithm for Extracting the Root of Arabic Words, 2009), and ARS(Arabic Light Stemmer (ARS), 2014).Some of these problems were solved by Algorithm2.A such as "قاهرة، محمد، ماعون، مرجان." In addition, other problems have been solved by Algorithm2.C. "مصائر، تمارين، مضامين،" And by Algorithm2.C Such as "مساهمات، مناسبة، التراثية ." Until the stemming algorithm that have adopted the rules or hybrid systems like Hadni(A new and efficient stemming technique for Arabic Text Categorization, 2012),Al-Kabi(A novel root based Arabic stemmer, 2015), Cherif(Building a syntactic rules-based stemmer to improve search effectiveness for arabic language, 2014), faced many problems that cannot be solved using these rules, but our approach solves it, such as " متلهفون ، لمحافظون ، دوامه فتيان ، الاطباء ، العلاج " by Algorithm2.C.

Another challenge is to choose prefixes and suffixes to be removed. However, removing some prefixes and suffixes may lead to trouble as it could adversely affect that eliminate an original word. For example, the word (اللهGod) if we delete the definite (الـ the), it becomes (له him) and then the word will be deleted as a preposition. To resolve these challenges, we built up a solution by collecting all exceptional words to exclude them from the removing process. Exceptional words that showed in (Table 9), is built by a lot of experiments in huge datasets. We believe this list can be used by other researchers to open new techniques and approaches in the Arabic text processing.

3.9. Arabic Text Categorization

In this study, we implement Arabic TC in two different approaches. The first approach is based on machine learning techniques (SVM, KNN, ANN, DT and NB), and the second approach is based on SVD. The following Figure describes our approaches.

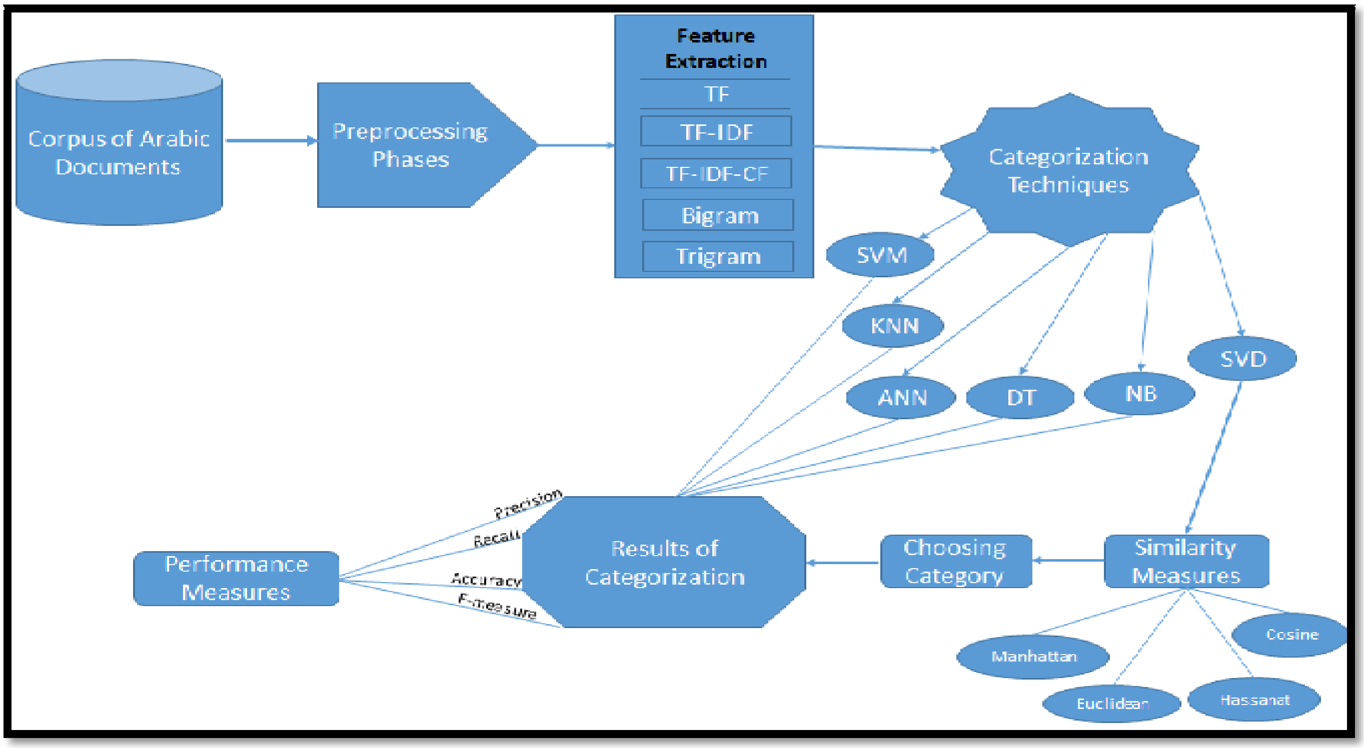


Figure 12
Our Framework's Steps

3.9.1 Arabic Text Categorization using Machine Learning

We implemented the machine learning techniques by using Waikato Environment for Knowledge Analysis (WEKA) tool (The WEKA data mining software: an update, 2009). It is software for data mining purposes. Weka is designed using Java programming language. It is widely used

machine learning algorithms (Research of text categorization on WEKA, 2013) for clustering, regression, categorization, preprocessing, visualization, etc. We use Weka 3.7.13 in our experiments for categorization.

"Supervised learning" is one of the appropriate types of machine learning which used for text categorization. When using machine learning, a training set and a test set are necessary. First, the training set is used for the classifier to learn the features that differentiate the documents. Then, a test set is used to confirm how well the automatic classifier performs. In this thesis, we employed five machine learning techniques, Naïve Bayesian (NB), Artificial Neural Networks, Decision Tree (DT), K-Nearest Neighbor (KNN) and Support Vector Machines (SVM). The following subsections describe these techniques.

The sake of clarity, we put an example to show the principle of the work for each of these techniques. Let's assume (T) distinct terms after preprocessing phase, the documents in training and testing are represented as t-dimensional vectors. See the following example:

D1 = 1T1 + 4T2 + 5T3
D2 = 6T1 + 0T2 + T3
test = 0T1 + 5T2 + 6T3

3.9.1.1 Real Example based on KNN

KNN learning algorithm calculates similarity between test document and all documents existing in the training set then returns test into the category which belongs to document that achieved the highest degree of similarity. In KNN, the Value of k is usually odd to avert neutralize, such as 1, 3 and 5, that most commonly used. In this study we use K=1.
Euclidean Distance(ED) is default with KNN classifier (KNN model-based approach in classification, 2003), chapter five illustrates ED in detail. According to previous example, ED is computed between test document D1 and D2 as follows:

ED(D1,Test)= $\sqrt[2]{(\mathbf{1}-\mathbf{0})^{2}+(\mathbf{5}-\mathbf{6})^{2}+(\mathbf{4}-\mathbf{5})^{2}}=\sqrt{\mathbf{3}}$ = 1.73

ED(D2,Test)= $\sqrt[2]{(\mathbf{6}-\mathbf{0})^{2}+(\mathbf{0}-\mathbf{5})^{2}+(\mathbf{1}-\mathbf{6})^{2}}=\sqrt{\mathbf{86}}$ = 9.27

Then test document resembles more the first document from the second document. When K=1 the test document assign into the category which belongs document 1.

3.9.1.2 Real Example based on NB

NB Represents the document D as a set of (Wi,TFi) pairs, where Wi is word and TFi is term frequency.

D1 = {(term1,1),(term2,1),(term3,0)} ∈ Category1(C1)
D2 = {(term1,1),(term2,0),(term3,1)} ∈ Category1(C2)
For example, document1 contains two terms T1and T2 but it not contains T3. See (Table 12).

Table 12

Naïve Bayes Example for Document Categorization

Document	T1	T2	T3	Category
D1	1	1	0	**C1**
D2	1	0	1	**C1**
D3	1	0	0	**C2**
D4	0	1	1	**C1**
D5	0	1	0	**C1**
D6	0	0	1	**C2**
testD	1	1	1	**???**

Now, is F(testD) ∈ C1 OR F(testD)∈ C2?
For each Category, this technique builds a Pr(testD |Category=C1) and Pr(testD |Category=C2), according to the following equation.

$\Rightarrow$ arg max Pr($C1$| testD) = arg max Pr(testD | C1) * Pr(C1)

$$\Pr(w_1,...,w_n \mid \text{Category} = C1) = \prod_{i=1}^{3} \Pr(w_i \mid \text{Category} = C1) \quad \text{……………………….(13)}$$

Pr(testD |Category=C1) which gives testD the best score according to Pr(testD |C1),pick word 1 according to Pr(term1|C1) then repeat for word 2,3. Each word is generated independently of the others as follows:

F(testD) = argmaxyPr(testD |C1)*Pr(C1).

Pr(Category=C1) = 1/2	Pr(Category=C2) = 1/2
Pr(T1=C1) = 1/2	Pr(T1=C2) = 1/2
Pr(T2=C1) = 3/4	Pr(T2=C2) = 0
Pr(T3=C1) = 1/2	Pr(T3=C2) = 1/2

Pr(T1,T2,T3| Category=**???**)
Pr(Category=C1)* Pr(T1=C1) * Pr(T2=C1)* Pr(T3=C1)
=0.5*0.5*0.75*0.5 = 0.09
Pr(Category=C2)* Pr(T1=C2) * Pr(T2=C2)* Pr(T3=C2) = 0
The maximum probability is with Category C1.

3.9.1.3 Real Example based on DT

Decision tree technique is used widely for text categorization. It achieves high accuracy results, so it is considered as one of the most effective algorithms. According to the previous table, we will apply the same example to find a category but using decision tree. We use WEKA with C4.5 by Ross Quinlan (Quinlan, 2014).

Figure 13 shows a tree for previous example where T represents term and C Category. In the test document, where T1, T2 and T3= 1, it belongs to Category1, follow red path.

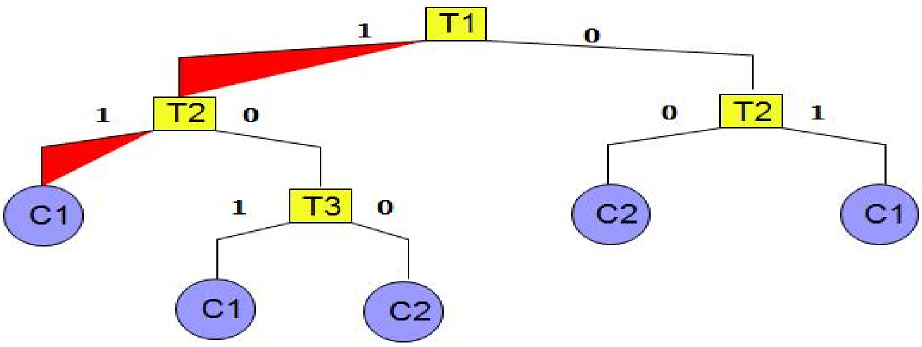


Figure 13
Decision Tree Example for Document Categorization

3.9.1.4 Real Example based on ANN

There are many neural network models; the perceptron is the oldest of these models. It was invented by (Rosenblatt, 1962). This model computes a weighted sum of inputs for generate the output depending on an activation function. There are many types of activation function such as; **Step** Function which based on 0/1- type function, if x > 1 then 1, otherwise then 0. Other function is **Sigmoid**; it is very useful in multi-layer networks. It is given according to the following equation:

Sigmoid function = 1/ (1+EXP(-input)) ……………….………….. (14)

In addition is **Hyperbolic tangent** formula which given by:

$\text{Tanh(input)} = \frac{\sinh(\text{input})}{\cosh(\text{input})} = (1-EXP(-2*input))/(1+EXP(-2*input))$……(15)

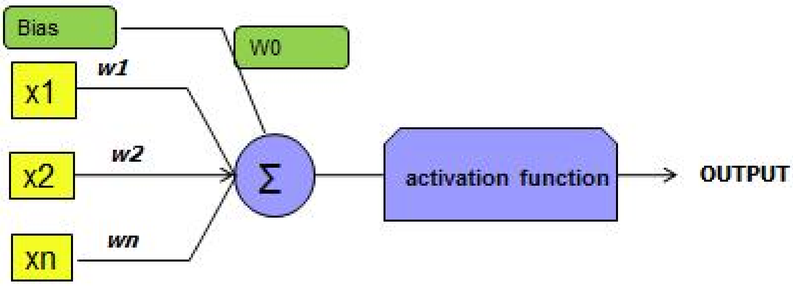


Figure 14
ANN Example for Document Categorization (Perceptron models)

In the first document suppose all the weights = 0.5 with x1 = 1, x2=1, x3=0, Threshold =1, Bias=1, where W is the vector of weights and B is the bias. According to previous example document1 is 1,1,0 and belongs to category C1 when actual output is 1 and class is C1, should increase the value of weights by little bit otherwise decrease weights by little bit and so on all document.

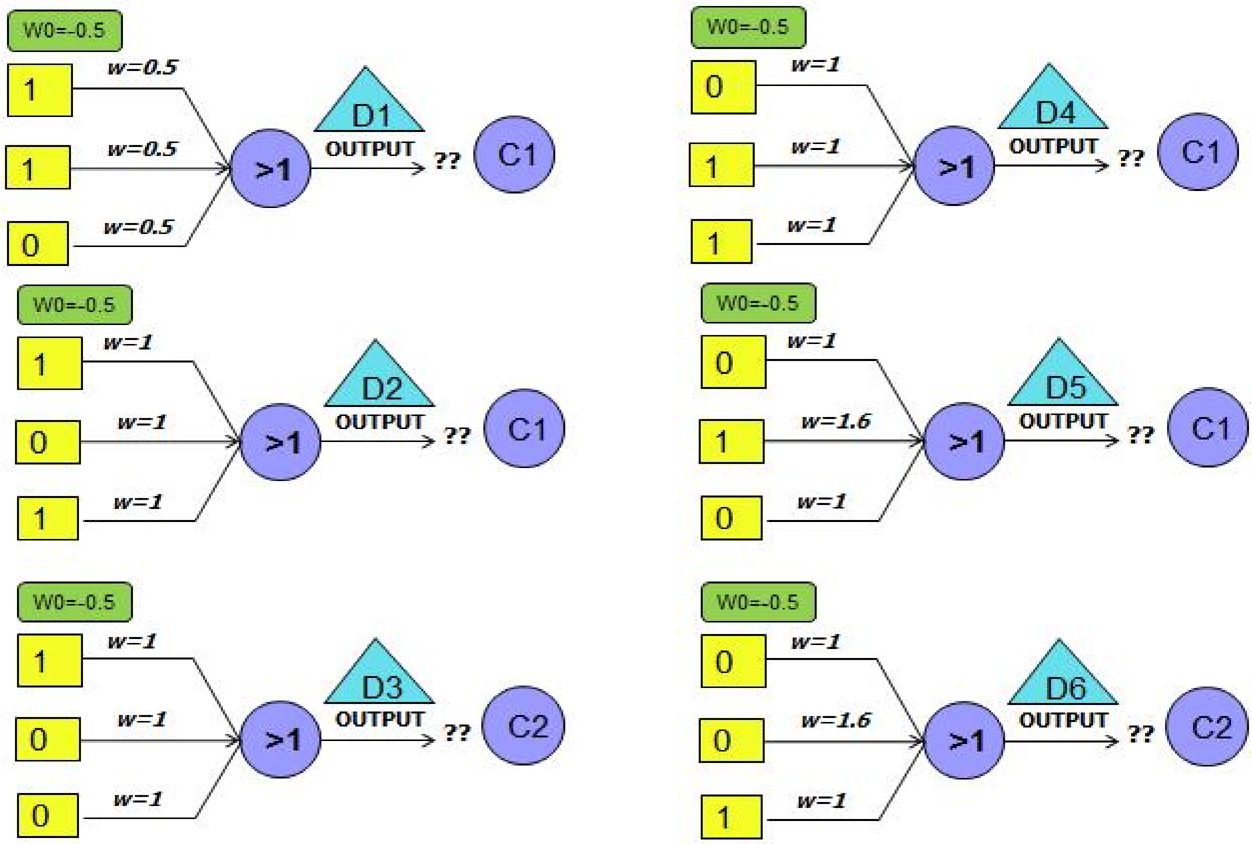


Figure 15
ANN Example for Document Categorization

In D1, D2, D3 and D4 all results are correct but D5 is wrong, so adjust the weights a little bit. Now for test document see (Figure 16).

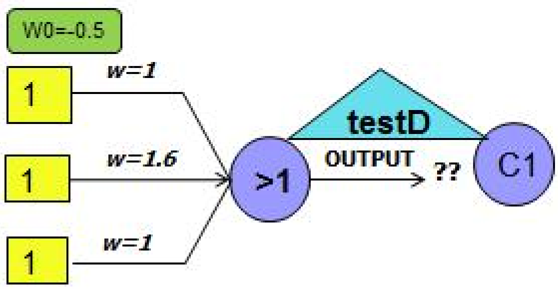


Figure 16
ANN Example for Document Categorization (Test Document)

Table 13 shows the result by step, Sigmoid and Hyperbolic tangent functions. From this table the text document belongs to category C1.

Table 13

Using Sigmoid and Hyperbolic tangent functions in ANN

Document	T1	T2	T3	Inputs	Threshold=1	Sigmoid	Tanh
D1	1	1	0	2.1	1	0.89	0.97
D2	1	0	1	1.5	1	0.82	0.91
D3	1	0	0	0.5	0	0.62	0.46
D4	0	1	1	2.1	1	0.89	0.97
D5	0	1	0	1.1	1	0.75	0.80
D6	0	0	1	0.5	0	0.62	0.46
testD	**1**	**1**	**1**	**3.1**	**1**	**0.96**	**1.00**

3.10 Arabic Text Classification Using SVD

In order to classify an Arabic text, the file should be represented as a vector. After that, a similarity function is used to find out what is the best category for the target file. The proposed approach used the SVD technique for file representation and the cosine similarity measure. This approach was taken because it gives the best results in comparison with other approaches, as will be shown later in details.

3.10.1 File Representations

Our approach of representing Arabic text is based on Latent Semantic Analysis (LSA). LSA is a sophisticated mathematical method that applies Singular Value Decomposition (SVD)(Indexing by latent semantic analysis, 1990). to the term-by-documents matrix to Determine the ‘‘latent semantic concepts’’ It is worth mentioning that SVD technology was designed for natural language processing applications and is commonly used in the field of information retrieval. There are few works that used SVD in the categorization of English texts, such as (Text classification by aggregation of SVD eigenvectors. In Advances in Databases and Information Systems, 2012). We did not find any research for Arabic documents categorization purposes using SVD. In our approach, each document is represented as a vector in the space of latent semantic.

SVD decompose the rectangular matrix A_{m*n} into three simpler matrices in order to extract the best features to express the big data and reducing the dimensionality (Information Systems Design and Intelligent Applications, 2015). It is worth mentioning the calculations made by SVD need a long time to execute, especially when there are a large number of terms in the document (Fast low-rank modifications of the thin singular value decomposition, 2006).

The rectangular matrix $A_{m\times n}$ can be split into three matrix components according to follow formula:

$$A_{m\times n} = U_{m\times m}\Sigma_{m\times n}V^{T}_{n\times n} \qquad (16)$$

The columns of $U_{m\times m}$ consist of orthonormal eigenvectors of AA^T, where $U.U^T$ = Identity matrix (I). (Recall from linear algebra that, orthogonal vectors of unit length are called 'orthonormal').

The values in Σ were formed from the diagonal elements of the matrix, which consists of square roots of eigenvalues of U or V (which are equal) those elements ordered descending. It expresses the linear difference of independent vectors along each dimension in a manner similar to the eigenvalues that express differences illustrated by 'factors' or vectors.
The columns of V are the transposes of an orthogonal matrix of V of size n x n. It describes orthonormal eigenvectors of $A^T A$, where $U^TU=I$, $V^TV=I$. To clean data and get rid of the noise, In order to reduce the dimensionality, we must delete some rows from the bottom of the matrices U and S and to keep the columns in the matrices S and V^T as it, shown in (Figure 17).

$$\begin{bmatrix} & & \\ & A & \\ & & m*n \end{bmatrix} = \begin{bmatrix} & & \\ & U & \\ & & m*m \end{bmatrix}\begin{bmatrix} & & \\ & \Sigma & \\ & & m*n \end{bmatrix}\begin{bmatrix} & & \\ & V & \\ & & n*n \end{bmatrix}$$

Figure 17

The Decomposition of Matrix A $_{m\times n}$ into U $_{m\times m}\Sigma_{m\times n}$ V $^{T}_{n\times n}$

SVD provides the mathematical foundation for text categorization and latent semantic indexing. Matrix A constructed through the intersection between terms and documents.

All the terms in the corpus are placed one after the other without repetition. As a result terms are represented as rows in the matrix A, while documents represent columns. The intersection of a column and a row represents the frequency of the term in the document.

There is a big challenge faced by this representation. This challenge is that matrix A will be very large and high-dimensionality. For the purpose of controlling this problem the use of the SVD is essential, where it reduces high-dimensional data by eliminating the noise and redundancy data so that data becomes more pronounced after eliminating terms that are repeated in almost all documents.

In all used datasets, singular value decomposition can be represented by matrix A (m* n) as a term-by document, where m are distinct terms in n documents. Always m is greater than n (m ≥n).

According to text categorization, previous equation, Formula number 12, of SVD is applied where:

$U_{m\times m}$ is column matrix of the document vectors

$\Sigma_{m\times n}$is diagonal matrix consisting of square roots of eigenvalues of U or V.

$V^T_{n\times n}$ is row matrix of the term vectors.

Each document represented in the form of the vectors such that vectors have the same length. Those vectors are covering all the unique words in all documents in the corpus, specifically in the training data set. Frequency of a particular term in a specific document represents the rate of term frequencies in frequency vectors. In some times, various studies used words instead of terms, the dimension values instead of term frequencies and the document vectors instead of frequency vectors.

In the following example, we present how vector space model is applied in the following documents:

1- اكد رئيس هيئة مفوضية المنطقة الاقتصادية الخاصة في العقبة بان عدد الطلبات المقدمة للاستثمار في المنطقة بلغت (47) طلبا.
2- نفى رئيس هيئة المفوضية وجود أي تجاوزات في التعيينات الاخيرة في المنطقة الاقتصادية الخاصة.
3- اذا لديك عشرة اجهزة حاسوب فكيف يمكنك اعداد شبكة وربطها من حاسوب الى حاسوب اخر ومن ثم ربطها بشبكة الانترنت .
4- عندما تقوم بربط شبكة حواسيب داخلية مع شبكة حواسيب خارجية تصبح لديك شبكة واسعة النطاق.
5- لم ينتهي حد التعاون الإسرائيلي الأوروبي عند هذا الحد ولكنه تجاوز إلى منح فرنسا أساسيات الطاقة النووية لإسرائيل، ثم ساهمت بريطانيا في إدخال هذه الطاقة في المجال العسكري.
6- لجأت فرنسا إلى التخفي وراء اتصالاتها التي أجرتها مع المقاومة العربية للتوصل إلى حل هادئ يرضي الجانب الإسرائيلي.
7- يواصل الفريق الأول بنادي الهلال السعودي تمارينه الاعتيادية استعداداً للقاء القادسية في دور الثمانية من مسابقة كأس ولي العهد.
8- قامت صحيفة الرياضية السعودية بمحاولة فاعلة لازالت الخلاف الذي نشب بين رئيس نادي الهلال الأمير محمد بن فيصل ومنصور البلوي في مقر الرياضة.

We will design any matrix from the terms that appear in more than one string, for the purposes of clarification.

The following example illustrates the idea using a small sample, we assume that there are four different categories: Computer, Economics, politic, and Sport. Eight documents in the training group and fourteen unique terms, as shown in (Table 14). Then, classify a new document based on the calculating of the similarity between the new document and the all

documents of the training group. After that, calculate similarities between test documents with training documents, as it will be explained later in the testing stage. Columns in (Table 14) represent the document's vector in every document that intersected with it, row represents the term frequencies. We chose four documents for each of the categories, D1and D2 from the Economics category, D3 and D4 from the Computer category, D5-D6 from the politic category and D7-D8 from the Sport category. In this example the length of the vector is 14, according to the number of unique terms. Through the table, we can note that the terms 1 and 2 represent the first Category because of the large amount of frequency to be in the documents that have been selected from this category. Thus, the terms 3 and 4 represent a second category and so on.

Table 14

Frequency Vectors of the Documents

Category	**Economics**		**Computer**		**Politic**		**Sport**		**Test**
TERM	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**	**D9**
منطقه	2	1	0	0	0	0	0	0	0
اقتصاد	1	1	0	0	0	0	0	0	0
خاص	1	1	0	0	0	0	0	0	0
هيئه	1	1	0	0	0	0	0	0	0
رئيس	1	1	0	0	0	0	0	1	0
مفوض	0	1	0	0	0	0	0	0	0
ربط	0	0	2	1	0	0	0	0	1
شبك	0	0	2	3	0	0	0	0	1
حاسوب	0	0	3	2	0	0	0	0	2
فرنسا	0	0	0	0	1	1	0	0	0
اسرائيل	0	0	0	0	2	1	0	0	0
نادى	0	0	0	0	0	0	1	1	0
هلال	0	0	0	0	0	0	1	1	0
رياض	0	0	0	0	0	0	0	1	0

It is clear from (Table 14), which represent the matrix A, that the number of terms is equal to 14, this represents the length of the vector. The frequency vectors contain frequencies of each term in each document.

3.10.2. SVD MODEL

Three matrices resulting from decomposition of the matrix A which are: U, Σ, V^T, The following steps represent calculation operations for each one of these three matrices.

First, calculate the matrix U:

U is the matrix of the eigenvectors of AA^T, it called document matrix size mxm' and this reflects the relationship of each document with the other documents, in other words, document-document matrix. There are five steps to calculate matrix U, for illustration, we will calculate U depending on the matrix A according to the previous example.

Step 1: compute AT
Step2: compute A^TA
Step3: compute eigenvalues of A^TA.
Step4: compute eigenvectors according previous eigenvalues.
Step5: Place the resulting eigenvectors together, so that the matrix U is formed. The resultant matrix U is shown in the following figure (figure 18)

0.00	-0.59	-0.00	-0.15	0.00	-0.56	0.02	0.00
0.00	-0.38	-0.00	-0.09	0.00	0.09	-0.14	-0.00
0.00	-0.38	0.00	-0.09	0.00	0.09	-0.14	-0.00
-0.00	-0.38	0.00	-0.09	0.00	0.09	-0.14	-0.00
-0.00	-0.42	0.00	0.28	-0.00	0.23	0.48	-0.00
0.00	-0.18	-0.00	-0.03	0.00	0.75	-0.30	0.00
-0.40	-0.00	0.00	-0.00	-0.50	-0.00	-0.00	0.00
-0.64	-0.00	0.00	0.00	0.75	0.00	0.00	-0.00
-0.66	-0.00	-0.00	-0.00	-0.43	-0.00	0.00	-0.00
-0.00	-0.00	-0.53	-0.00	0.00	0.00	0.00	0.85
-0.00	-0.00	-0.85	-0.00	-0.00	0.00	-0.00	-0.53
0.00	-0.05	-0.00	0.61	0.00	-0.12	-0.34	0.00
0.00	-0.05	-0.00	0.61	0.00	-0.12	-0.34	0.00
0.00	-0.04	-0.00	0.37	-0.00	0.14	0.62	0.00

Figure 18
Matrix U (Document Matrix)

Second, calculate the matrix V:
V is the matrix of the eigenvectors of A^TA, it called term matrix of size nxn, and this reflects the relationship of each term with the other term, in other words, term-term matrix. There are four steps to calculate matrix V, a according to the previous example.
Step1: compute AA^T
Step2: compute eigenvalues of AA^T.
Step3: compute eigenvectors according to previous eigenvalues.
Step4: Place the resulting eigenvectors together, so that the matrix V is formed.

0.00	0.00	-0.74	-0.67	-0.00	-0.00	-0.00	-0.00
-0.75	-0.64	0.00	-0.00	-0.00	-0.00	-0.03	-0.16
0.00	-0.00	0.00	-0.00	-0.85	-0.53	0.00	-0.00
-0.13	-0.08	-0.00	0.00	-0.00	-0.00	0.54	0.83
0.00	-0.00	-0.67	0.74	-0.00	0.00	0.00	-0.00
-0.63	0.72	-0.00	0.00	-0.00	0.00	-0.25	0.13
0.14	-0.25	-0.00	0.00	-0.00	0.00	-0.80	0.52
-0.00	0.00	0.00	0.00	-0.53	0.85	0.00	-0.00

Figure 19
Term Matrix V^T

Thirdly, calculate the matrix Σ:
Step1: Compute the square roots of eigenvalues from U or V.
Step2: Order these values descending.
Step3: Place diagonally in the matrix and fill the rest of the matrix zeros, so that the matrix Σ is formed.

5.44	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	3.65	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	2.62	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	2.25	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	1.19	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.97	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.84	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.38

Figure 20
Matrix Σ (diagonally)

For calculating SVD samples, we can use Blue Bit calculator tool.After applying the SVD, we can classify any new document, suppose we want to test a new document, we must first create a new vector of this document, so that it is equal to a length of 16 for each unique term by training group. To illustrate the idea, consider the following vector expresses the unique terms frequencies in the new document.

"شبكة الحاسوب هي ربط اكثر من جهاز حاسوب معا"

NewDocumentTest = [0,0,0,0,0,0,1,1,2,0,0,0,0,0]

Evidenced by the number of frequencies that the new document is closer second category. Which means that this document is selected from the second category. Actually, it does not matter the number of frequencies in the new document, what is important is only the appearance of the term in the document or not. Thus, new document can be represented as follows:
NewDocumentTest_prime =[0,0,0,0,0,0,1,1,1,0,0,0,0,0]

It is noted through the previous matrix Σ that there are very small values. These values are considered noise and adversely affect the accuracy of the categorization, where increases the burden computational. For these reasons, we will use the SVD with K values.

These values must be removed to reduce dimensionality, this is done through replace all these negligible values by zeros, and this means that a certain amount of the singular values will be maintained. Let us assume that this amount represents K. Thus, the entire matrix becomes zeros except the Initial k values along its diagonal.

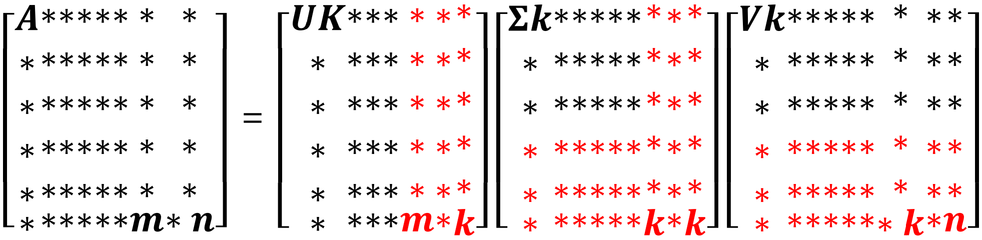


Figure 21
SVD after K Reduction

Matrix Σ reduced to Σ_K in addition to U and V_T reduced to U_k and $V^T{}_k$ respectively. The main equation that represents the relationship between terms and documents will become as follows

$A_k = U_{m*k}\,\Sigma_{k*k}\,V^T{}_{k*n}$……………………………………………………… (17)

Where:
U_k is the first k rows of the matrix U, $U_k\Sigma_k$ represented the terms.
$V_k{}^T$ is the first k columns of matrix VT, $V^T{}_k\Sigma_k$ represent the documents.
Σ_k is a diagonal matrix presented the initial k factors.
Follow-up to the previous example. Let K=5,
U5:

0.00	-0.59	-0.00	-0.15	0.00
0.00	-0.38	-0.00	-0.09	0.00
0.00	-0.38	0.00	-0.09	0.00
-0.00	-0.38	0.00	-0.09	0.00
-0.00	-0.42	0.00	0.28	-0.00
0.00	-0.18	-0.00	-0.03	0.00
-0.40	-0.00	0.00	-0.00	-0.50
-0.64	-0.00	0.00	0.00	0.75
-0.66	-0.00	-0.00	-0.00	-0.43
-0.00	-0.00	-0.53	-0.00	0.00
-0.00	-0.00	-0.85	-0.00	-0.00
0.00	-0.05	-0.00	0.61	0.00
0.00	-0.05	-0.00	0.61	0.00
0.00	-0.04	-0.00	0.37	-0.00

Figure 22
Matrix U where K=5

Σ5:

5.44	0.00	0.00	0.00	0.00
0.00	3.65	0.00	0.00	0.00
0.00	0.00	2.62	0.00	0.00
0.00	0.00	0.00	2.25	0.00
0.00	0.00	0.00	0.00	1.19

Figure 23
Matrix Σ5, where K=5

V^T5

0.00	0.00	-0.74	-0.67	-0.00	-0.00	-0.00	-0.00
-0.75	-0.64	0.00	-0.00	-0.00	-0.00	-0.03	-0.16
0.00	-0.00	0.00	-0.00	-0.85	-0.53	0.00	-0.00
-0.13	-0.08	-0.00	0.00	-0.00	-0.00	0.54	0.83
0.00	-0.00	-0.67	0.74	-0.00	0.00	0.00	-0.00

Figure 24
Matrix V^T5, where K=5.

The next step calculates term vector by $U_5*\Sigma_5$ to terms 7, 8 and 9 because it appears in the new test document. Then documents vector multiplied by $V^T_{5*}\Sigma_5$ to all documents.

After multiplication by Σ, documents from 1 to 8 become as follows:

Table 15
Document Representation

D1	D2	D3	D4	D5	D6	D7	D8
0	0	-4.05	-3.64	0	0	0	0
2.74	-2.33	0	0	0	0	0.1	-0.57
0	0	0	0	-2.23	-1.38	0	0
-0.28	-0.17	0	0	0	0	1.22	1.68
0	0	-0.8	O.98	0	0	0	0

Now the new test document is computed as the centroid of the vectors terms appeared. Now, represent the new document as a vector:
[-4.05, 0, 0, 0, -0.8]
ND-test = [-0.43,-0, 0, 0,-0.52]

Thus, for all the remaining documents according matrix A, The Table Then compare this vector with 8 vectors or 8 rows in matrix U and calculates the amount of similarities by similarity measures.

3.10.3. SIMILARITY MEASURES

There are many of metrics used to solve problems of natural language processing (Similarity measures for text document clustering, 2008)(Vector Space Models to Classify Arabic Text., 2014), such as Euclidean distance, Cosine distance, Jaccard, Dice. Similarity or distance measures can be used to express the degree of closeness between letters, words, strings or documents.

There is no measure that gives the best results for all kinds of problems. The decision is related to the word processing (Similarity measures for text document clustering, 2008)due to the nature of the data

and the type of problem. Therefore, the choice of an appropriate measure of similarity is essential.

For the purposes of comparison vectors together and measuring the proportion of similarities between them, there are several ways. In this thesis we used Euclidean, Cosine, and Hassanat's distances.

3.10.3.1. Euclidean Distance Metric

This distance is also called as the "Pythagorean theorem".(Short Text Classification Using kNN Based on Distance Function, 2013) It is a common metric to calculate distance between the vectors, and widely used in text categorization field, especially when using the KNN algorithm. In this distance, the vector is represented as a single point, then calculates the distance between them by the square root of the total sum of the differences between those two points box. According to the following equation:

$$\text{ED (a,b)}= \sqrt[2]{\sum_{i=0}^{n-1}(\mathbf{a_i}-\mathbf{b_i})^2} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots. (18)$$

Where:

a and b: The frequencies of terms in training and testing document.

For example,

ND-test = [-0.43,0, 0, 0,-0.52]

Document3= [-4.05, 0, 0, 0, -0.8]

$=\sqrt{(-3.62)^2 + (-0.28)^2}$

$= \sqrt{13.10 + 0.07}$ =**3.63**

3.10.3.2. Cosine Similarity Metric

Cosine similarity is a metric used to compute amount of similarity between two vectors, according to following formula:

$$\text{Cosine Similarity (Dki,Tk)} = \frac{\sum_{k=1}^{n} Dki*Tk}{\sqrt{\sum_{k=1}^{n} D^2ki} * \sqrt{\sum_{k=1}^{n} T^2ki}} \dots\dots\dots\dots\dots\dots\dots\dots\dots.(19)$$

Where:

D_{ki}: The frequency of term k in the document i.

T_k: The frequency of term k in the new document test.

Similarity (D_{KI},T_k) :the cosine similarity of D_{ki} and T_k.

We can also define it as the cosine of the angle between D_{ki} and T_k. (Figure 25) shows the similarity between test vectors colored in blue with D3 and D8, Observe that the test vector is similar to document 3, (d3).

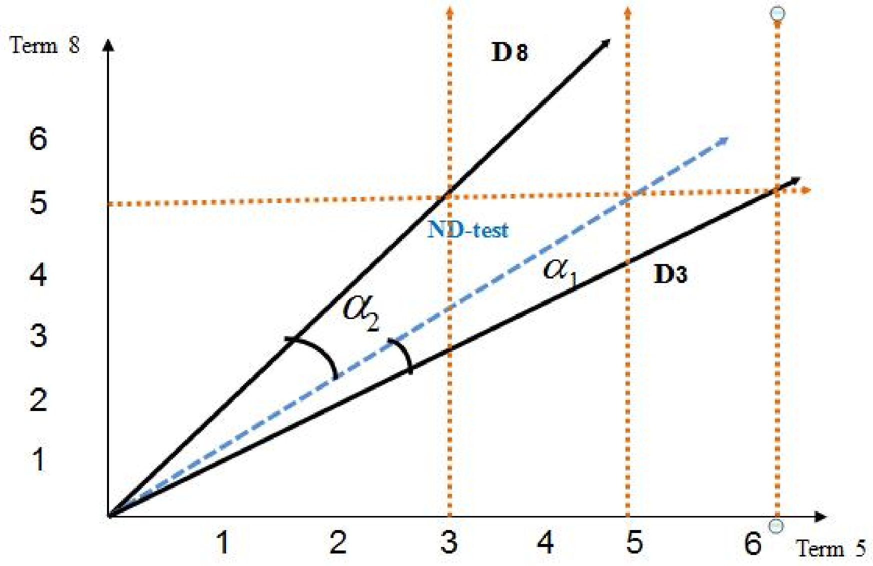


Figure 25
Cosine Distance between two Terms (5 and8)

For example,
ND-test = [-0.43,0, 0, 0,-0.52]
Document3= [-4.05, 0, 0, 0, -0.8]
Cosine Similarity (ND-test, d3) = dot(ND-test, d3) / ||ND-test|| ||d3||
dot(ND-test, d3) = (-0.43)*(-4.05) + (0)*(0) + (0)*(0) + (0)*(0) + (-0.52)*(-0.8) = **2.1575**
||ND-test|| = sqrt((-0.43)^2 + (0)^2 + (0)^2 + (0)^2 + (-0.52)^2) **= 0.674759216313**
||d3|| = sqrt((-4.05)^2 + (0)^2 + (0)^2 + (0)^2 + (-0.8)^2) = **4.12825629049**
Cosine Similarity (ND-test, d3) = 2.1575 / (0.674759216313) * (4.12825629049)
= 2.1575 / 2.78557897931

$$\approx \mathbf{0.77}$$

3.10.3.3. Hassanat's Distance Metric (HD)

HD is a distance method, which used to compute the distance between two vectors(Dimensionality Invariant Similarity Measure, 2014), this method proposed a solution for the treatment of the problem of dispersion in dimensions between Vector. Highlights the importance of this method when there is significant variation in the outliers. There is no needing for data normalization before categorize it. It is computed according to the following formula:

$$D(A_i, B_i)\begin{cases} 1 - \dfrac{1 + \min(Ai, Bi)}{1 + \max(Ai, Bi)} & \min(Ai, Bi) \geq 0 \\ 1 - \dfrac{1 + \min(Ai, Bi) + |\min(Ai, Bi)|}{1 + \max(Ai, Bi) + |\min(Ai, Bi)|} & \min(Ai, Bi) < 0 \end{cases} \ldots\ldots\ldots\ldots.(20)$$

$HD(A, B) = \sum_{i=1}^{n} D(A_i, B_i)$..(21)

Where:

A , B: The frequencies of terms in training and testing document.

$D(A, B)$: The distance between each two points in vectors.

$HD(A, B)$: The sum of distances between all points in the vectors.

The following example illustrates the efficiency of Hassanat distance when there is a particular term of high-frequency in one of document and a little frequency in the test text.

ND-test = [-0.43,0, 0, 0,-0.52]

Document3= [-4.05, 0, 0, 0, -0.8]

$$HD = \left(1 - \frac{1 + \min(-0.43, -4.05) + |\min(-0.43, -4.05)|}{1 + \max(-0.43, -4.05) + |\min(-0.43, -4.05)|}\right) + \left(1 - \frac{1 + \min(-0.52, -0.8) + |\min(-0.52, -0.8)|}{1 + \max(-0.52, -0.8) + |\min(-0.52, -0.8)|}\right)$$

$$HD = \left(1 - \frac{-3.05 + 4.05}{0.57 + 4.05}\right) + \left(1 - \frac{0.2 + 0.8}{0.48 + 0.8}\right)$$

$$HD \to (0.78) + (0.22) = \mathbf{1}$$

3.10.3.4. Manhattan Distance (MD)

As for the distance of Manhattan, also called Taxicab geometry, (Taxicab Geometry: An Adventure in Non-Euclidean Geometry., 1986) (Implementing Discrete Mathematics: Combinatorics and Graph Theory with Mathematica, 1990) It is used to calculate lengths of paths between points that represent curves in the XY axes, this idea presented in the nineteenth century by replacing the Euclidean geometry. So that the distance between two points is the sum of the absolute values of differences and Coordinates. The most common example of this method is taxi between cities, so the distance traveled by car will not differ. It is given according to the following equation:

$MD(A, B) = \sum_{i=1}^{n} |A_i - B_i|$ .... (22)

Where:

A and B: Are the frequencies of terms in training and testing document.

$MD(A, B)$: Is the distance between each two points in vectors.

According to the previous example, MD between ND-test and Document3 can be computed as follows:

ND-test = [-0.43, 0, 0, 0,-0.52] , Document3= [-4.05, 0, 0, 0, -0.8]

$$MD = |-0.43 + 4.05| + |-0.52 + 0.8| = 3.62 + 0.28 = \mathbf{3.9}$$

You should pay attention that both Euclidean, Hassanat and Manhattan are distances not similarity measures. So what concerns us here, the file that achieves the lowest value with the test document, so, that it is affiliated to the same category. Dislike, cosine similarity.

3.10.4. Categorizing file Choosing

Now, let's apply the similarity metric on the previous example. For calculating similarity on small examples, we can use cosine similarity calculator tool. At (2012)

Results in (Table 16) shows that the 3rd document achieves the highest percentage of similarity, but actually we cannot consider the test document in the same category that contains the document that achieves the highest similarity with the test document. To overcome this problem, some studies such as (Limits on the majority vote accuracy in classifier fusion, 2003) used the majority vote, in this study, they chose the best of three and five then they vote to category, finally, they chose the category to which it belongs with the highest percentage of similarity. In our example, we use majority vote with 2, documents that have achieved the highest similarity ratios are documents 3 and 4, the new document is classified as belonging to the second category, this result is correct. Although the results that appeared were correct results but in fact there are problems facing this way, these problems that real datasets are huge, where the matrix has a very large number of unique terms and documents.

Table 16

Similarity between Test Documents (ND-test) with the 8 Documents (after eliminating noise)

Documents	Hassanat's distance with test document	Euclidean distance with test document	Cosine distance with test document	Category average based on Cosine Similarity
D1	0.68	0.74	0.00	0.00
D2	0.70	0.71	0.00	
D3	**0.80**	**0.78**	**0.78**	0.61
D4	0.73	**0.78**	0.44	
D5	0.73	0.70	0.00	0.00
D6	0.76	0.61	0.00	
D7	0.74	0.58	0.00	0.00
D8	0.67	0.67	0.00	

Chapter Four:
Experiments and Results

4. Experiments and Results
4.1. Used Data Sets

Most of the Arabic studies on the categorization of Arabic texts built their data set from online web sites. For training, testing and evaluation of the results of their method. So that, the results of each study was based on in-house data set. Through deep research with regard to the Arabic language, No standard Arabic data set available for free (Comparative evaluation of text classification techniques using a large diverse Arabic dataset, 2013)unlike English language, for the mentioned reasons, we have to communicate with many researchers to obtain the data set they used in their studies.

We have noticed that there is a problem in most of these data sets where it is equal in terms of the number of documents or text files in each category, which affects the accuracy positively and negatively based on the number of documents in each category. For example, if we assume we have a category contains 50 documents and other category contains 250 documents, and let's say that all the documents contain an equal number of the distinctive terms, In this case the result will not be impartial and equitable, because the system will categorize test files based on number of documents which trains them, Then we can say that the number of terms for the first category, which contains 250 documents is increased by more than five times the second category, which contains 50 documents. So the accuracy of the system will be biased to the first category. To resolve this issue we built an identical data set containing the same number of documents in each category as we will address in details later.

We have obtained 11 different data sets, below is a description of this data set in terms of approach that is used for categorization, Evaluation metrics, the names of the categories, the number of documents and a number of terms in each category. For ease of dealing with these data sets, we have the adoption of a unified labeling, this label taken from the first letter following words Arabic Text Categorization Data Set "ATC-DS" then Number data set starts from 1 to 11.

ATC-DS1: This idiomatic Name to express Data set published in a year 2007 by Abdelwadood MESLEH, he have constructed a data set from five websites of Egyptian newspapers are Al-hayat, Al-Dostor, Al-Ahram Al-Nahar, and Al-Jazeera. This data sets contains a 1445 document distributed on nine categories. each category contains a set of documents as follows Computer with 70 documents, Economics with 220 documents, Education with 68 documents, Engineer with 115 documents, Law with 97

documents, Medicine with 232 documents, Politics with 184 documents, Religion with 227 documents, and Sports with 232 documents. He used three classifiers: SVMs Classifier, Naïve Bayes classifier, k-NN classifier. This classifier has achieved the following results respectively 88.11, 84.54, 72.72. (Chi square feature extraction based SVMs Arabic language text categorization system, 2007).

ATC-DS2: This idiomatic Name to express Data set called (Corpus of Contemporary Arabic (CCA). It published in a year 2013 by Hadni et al; they have constructed a data set from a large Arabic corpus. This data sets contains a 276 document distributed on 12 categories. each category contains a set of documents as follows economics with 29 documents, Education with 10 documents, health and medicine with 32 documents, interview with 23 documents, politics with 10 documents, recipes with 9 documents, Religion with 19 documents, science with 45 documents, spoken with 5 documents, Sociology with 30 documents, tourist and travel with 60 documents, and Sports with 4 documents. They used five Techniques of Feature Selection document frequency (DF), Information Gain (IG), Mutual Information (MI), Chi-Square (CHI) and CHIR.. All of them with Naïve Bayes classifier, CHI and CHIR achieved the superior results. (A Comparative Study with Different Feature Selection for Arabic Text Categorization)

ATC-DS3: This idiomatic Name indicates to Data set an in-house collected corpus from seven of the websites of Jordanian newspapers. As we have noted, data sets from ATC-DS1 to ATC-DS10 are unbalanced. For this purpose we built balanced dataset. It contains 880 documents with 9,183,511 words distributed on 11 categories, each category contains 80 documents, and these categories are Economy, Sports, Ideas and attitudes, Local News, health and medicine, Technology, Art, Education and Universities, Culture, Wars and Politics and Islamic. Each document was saved in a separate file within the corresponding category's directory. Our data base consists from two parts, the first part is from others, and the second part of data were collected at the period from January to October of 2015 from the following Jordanian newspapers: Alghad (http://www.alghad.com/), Alrai(www.alrai.com), Alanbatnews (http://www.alanbatnews.net,), Alarabiawm(http://www.Alarabiawm.net), Addustour(http://www.addustour.com), Assabeel(http://www.assabeel.net), hashd-ahali(http://www.hashd-ahali.net/main/ahali).

ATC-DS4: This idiomatic Name to express Data set published in a year 2006 by Syiam et al, they have constructed a data set from three websites of Egyptian newspapers are: El Gomhoria, El Akhbar, and El Ahram from August 1998 to September 2004.This data sets contains a 1,132 documents and with39,468 word distributed on six categories. each category contains a set of documents as follows art with 233 documents, Economics with 233

documents, Woman with 121 documents, Information Technology with 102 documents, Politics with 280 documents, and Sports with 231. 117 word was approximate average number of words in each document. They used two classifiers: k-NN classifier and the Rocchio classifier. Rocchio classifier has achieved results better than K-NN. (Arabic Text Classification Using Support Vector Machines, 2009),(An intelligent system for Arabic text categorization, 2006).

ATC-DS5: This idiomatic Name to express Data set called (Essex Arabic Summaries Corpus (EASC). This corpus published in a year 2013. This data sets contains a 153 document distributed on 10 categories with two encoding formats UTF-8 and ISO-8859-6 (Arabic). each category contains a set of documents as follows Art and Music with 9 documents, Education with 6 documents, health with 16 documents, environment with 32 documents, politics with 20 documents, finance with 16 documents, Religion with 7 documents, Science and Technology with 7 documents, sport with 9 documents and tourisms with 9 documents. It is available on (2015)

ATC-DS6: This idiomatic Name indicate to Data set called (Khaleej-2004 corpus), this corpus used in three studies (Comparison of topic identification methods for arabic language, 2005),(Multi-category support vector machines for identifying Arabic topics, 2009),(Topic identification by statistical methods for arabic language, 2006).It constructed from the Arabic newspaper “Alkhabar”. These data sets contain a 5690 document with 2.855.069 terms distributed on 4 categories. With each category contains a set of documents as follows: Local news with 2398 documents, Economy with 909 documents, International news with 953 documents, and sport with 1430 documents. It available on (2015)

ATC-DS7: This idiomatic Name indicate to Data set called (Watan-2004 corpus), this corpus used in five studies (Comparing TR-Classifier and KNN by using Reduced Sizes of Vocabularies, 2009),(Evaluation of Topic Identification Methods for Arabic Texts and their Combination by using a Corpus Extracted from the Omani Newspaper Alwatan, 2011),(Evaluation of Topic Identification Methods on Arabic Corpora, 2011),(Tr-classifier and knn evaluation for topic identification tasks, 2010).It constructed from the Arabic newspaper “Watan”. These data sets contain a 20291document with 9.813.366 terms distributed on 6 categories. with each category contains a set of documents as follows: Culture news with 2782 documents, Religion news with 3860 documents, Local news with 3596 documents, Economy with 3468 documents, International news with 2035 documents, and sport with 4550 documents. It available on(2015).

ATC-DS8: This idiomatic Name indicate to Data set called (The Akhbar-Alkhaleej Arabic), this corpus used in two studies (Feature subset selection for Arabic document categorization using BPSO-KNN, 2011),(Arabic text

categorization via binary particle swarm optimization and support vector machines, 2011) It constructed from the online Arabic newspaper “Alkhabar". These data sets contain a 5690 document distributed on 4 categories. With each category contains a set of documents as follows: Local with 576 documents, Economy with 218 documents, International news with 228 documents, and sport with 343 documents. It is available on (Khalifa).

ATC-DS9: This idiomatic Name indicate to Data set called (The Alwatan Arabic), this corpus used in two studies (Feature subset selection for Arabic document categorization using BPSO-KNN, 2011),(Arabic text categorization via binary particle swarm optimization and support vector machines, 2011) It constructed from the online Arabic newspaper "Alwatan". These data sets contain a 20,291document distributed on four categories. With each category contains a set of documents as follows: Culture with 156 documents, Economy with 194 documents, Religion with 216 documents, and sport with 255 documents. It is available on (Khalifa).

ATC-DS10: This idiomatic Name indicate to Data set called (The Al-jazeera-News Arabic), this corpus used in two studies (Arabic text categorization via binary particle swarm optimization and support vector machines, 2011),(Feature subset selection for Arabic document categorization using BPSO-KNN, 2011). It constructed from the online Arabic newspaper "The Al-jazeera". This data sets contains a 1500 documents distributed on five categories. With each category contains a set of documents as follows: Art with 240 documents, Science with 240 documents, Economy with 240 documents, politics with 240 documents, and sport with 240 documents. It is available on (Khalifa).

ATC-DS11: This idiomatic Name indicates to Data set called Corpus of Contemporary Arabic (CCA), this corpus contains 415 documents distributed on 14 categories. With each category contains a set of documents as follows: Autobiography with 73 documents, Children's Stories with 27 documents, Economy with 29 documents, Education with 10 documents, and sport with 4 documents. Health and Medicine with 32 documents, Interviews with 23 documents, Politics with 29 documents, Recipes with 9 documents, Religion with 19 documents. Science with 45 documents, Tourist and Travel with 60 documents, Short Stories with 31 documents Sociology with 30 documents. It available on (AL Sulaiti, 2009).

Table 17
Used Dataset Summary

Name of Database	Source	# of documents	# of Categories
ATC-DS1	five websites of Egyptian newspapers	1445	Nine
ATC-DS2	Corpus of Contemporary Arabic (CCA)	276	Twelve
ATC-DS3	seven websites of Jordanian newspapers and some of data sets	880	Eleven
ATC-DS4	three websites of Egyptian newspaper	1,132	Six
ATC-DS5	Essex Arabic Summaries Corpus (EASC)	153	Ten
ATC-DS6*	Arabic newspaper “Alkhabar”.	5690	Four
ATC-DS7*	Arabic newspaper “Watan”	20291	Six
ATC-DS8*	Arabic newspaper “Alkhabar"	5690	Four
ATC-DS9*	online Arabic newspaper "Alwatan"	20291	Four
ATC-DS10	online Arabic newspaper " The Al-jazeera"	1500	Five
ATC-DS11	Corpus of Contemporary Arabic (CCA)	415	Fourteen

*We chose the first 100 text file only from each category.

4.2. Experiments' setup

Our experiments are divided into two main parts. The first part is set to verify our Arabic Stemmer. In these experiments, we verify the accuracy improvements after using our proposed Arabic texts preprocessing. The second part is set to compare Arabic text classifications, which include machine learning techniques, as well as, SVD approach. These experiments are set to find out what is the most accurate techniques for Arabic texts. The following subsections show these two parts of the experiments. All of our experiments is done in a fixed specifications that are shown in (Table 18

Table 18

Our Experiments' Setup

Hardware setup	Computer Name	HP-PC
	Processor	Intel (R) Core i5-2450M CPU @ 2.50 GHz
	RAM	6.00 GB
	Hard disk	600 GB
Software Setup	Operating System	Windows 7 (64- bit)
	IDE	Eclipse for Java
	Programming Language	Java SE Development Kit 7u79
	Tool	Weka 3.7.13

4.3. Evaluation for the Proposed Arabic Stemmer

We chose SVM classifier to test the impact of our stemmer on the classifier. (Table 19) shows the accuracy measures of using SVM classifier with our Stemmer and with other well-known stemmer. As shown in Table 19, our stemmer has the highest accuracy.

Table 19

Comparison between Stemmers using SVM Classifier

Name of Database	Name of stemmer												
	Light 1	**Light 2**	**Light 3**	**Light 8**	**Light 10**	**P-Stemmer**	**Alstem**	**Umass**	**Modified Umass**	**Aljlayl**	**Mustafa**	**Taghva**	**Our Stemmer**
ACT-DS1	92%	92%	92%	92%	92%	92%	91%	92%	92%	92%	92%	92%	93%
ACT-DS2	73%	75%	73%	72%	75%	73%	76%	72%	76%	75%	73%	73%	76%
ACT-DS3	90%	90%	91%	91%	90%	90%	91%	91%	90%	90%	90%	91%	91%
ACT-DS4	86%	86%	86%	86%	86%	86%	85%	86%	86%	88%	87%	86%	88%
ACT-DS5	48%	48%	50%	50%	50%	48%	46%	54%	54%	42%	54%	48%	56%
ACT-DS6	93%	93%	94%	93%	93%	93%	94%	94%	93%	93%	92%	94%	94%
ACT-DS7	97%	96%	97%	97%	97%	96%	98%	97%	97%	97%	98%	97%	98%
ACT-DS8	92%	92%	92%	93%	92%	92%	93%	93%	93%	91%	91%	92%	93%
ACT-DS9	90%	90%	91%	90%	90%	90%	90%	90%	90%	95%	90%	98%	98%
ACT-DS10	96%	96%	96%	96%	95%	96%	96%	96%	96%	96%	96%	95%	96%
ACT-DS11	75%	75%	76%	75%	76%	75%	77%	76%	77%	75%	76%	76%	77%
Average	**85%**	**85%**	**85%**	**85%**	**85%**	**85%**	**85%**	**85%**	**86%**	**85%**	**85%**	**86%**	**87%**

Returns to the real root cause cancellation of indication the word class or category to which they belong, for example, the (جامعة University) (جامعmosque) and the entire (مجتمع community) goes back to the root (collect جمع) where you can trace the word (جامعة University) education class and the word (جامعMosque) to the category of religions and the word (مجتمع community) of class cultures, that had to take this issue into account

and try to trim the stem as much as possible without impact on the classification, for example (اقتصادeconomy) and (اقتصاديeconomic) and (اقتصادياتEconomics) and (الاقتصاديونeconomists) refers to one word (اقتصادeconomy) and the term becomes a repeating of (اقتصاد economy) four times.

But there was a clear challenge to choose prefixes and suffixes, to yield positive results as mentioned in (Table 19) but the removing of them may lead to trouble as it could adversely affect that eliminates an original word, for example, the word (الله، God) if we delete the definite (الـ the), it becomes (له, him) and then the word will be deleted as a preposition. To resolve these challenges, our solution to this problem without negative impact on the results. Where we have provided a solution for the excluded words (exceptional table) for all original words, and then convert all weights with alone to generate a term with the same meaning to raise the number of frequencies of the real term itself, and in the end we use the idea of stemming to remove prefixes and suffixes.

4.4. Using the Proposed Arabic Stemmer with Classifiers

4.4.1. KNN classifier based on the Proposed Stemmer

From (Table 20), we can conclude that KNN classifier is not suitable for document categorization; this is Due to a variety of reasons, some of which goes to the nature of the data, and others of the nature classifier, from the figure (Figure 8) note that the selection of the appropriate value of k is not easy.

Table 20

KNN with Our Stemmer

Name of Database	Testing Time (sec.)	Accuracy %	Precision %	Recall %	F-measure %
ACT-DS1	330	54%	81%	54%	58%
ACT-DS2	0.45s	35%	61%	35%	28%
ACT-DS3	1078	16%	82%	16%	10%
ACT-DS4	78	38%	58%	38%	39%
ACT-DS5	16	37%	33%	37%	29%
ACT-DS6	3940	75%	80%	75%	73%
ACT-DS7	147	78%	85%	78%	77%
ACT-DS8	410	71%	78%	71%	69%
ACT-DS9	6953	77%	84%	77%	77%
ACT-DS10	226	41%	82%	41%	39%
ACT-DS11	94	43%	61%	43%	39%
Average	**1327**	**51%**	**71%**	**51%**	**49%**

4.4.2: ANN Classifier Based on the Proposed Stemmer

ANN results as seen in (Table 21), proves that ANN is not a good classifier for Arabic documents. For example data set ACT-DS2 and ACT-DS3 Classification accuracy was not accepted (0.40 and 0.41)

Table 21

ANN with Our Stemmer

Name of Database	Testing Time	Accuracy %	Precision %	Recall %	F-measure %
ACT-DS1	237	65%	80%	65%	65%
ACT-DS2	155	41%	56%	41%	42%
ACT-DS3	281	41%	86%	41%	42%
ACT-DS4	409	53%	51%	53%	52%
ACT-DS5	200	39%	38%	39%	38%
ACT-DS6	114	83%	78%	83%	83%
ACT-DS7	109	82%	81%	82%	82%
ACT-DS8	2734	76%	79%	76%	76%
ACT-DS9	86763	85%	81%	85%	85%
ACT-DS10	209	61%	77%	61%	61%
ACT-DS11	123	46%	56%	46%	45%
Average	**8300**	**61%**	**69%**	**61%**	**61%**

4.4.3. Decision tree classifier based on the Proposed Stemmer

Decision tree classification accuracy ranged between 0.40 for ACT-DS5 and 0.91 For ACT-DS9. This shows that this method is appropriate for the classification of huge data, and does not fit the data with a few files such as ACT-DS5.

Table22

Decision Tree with Our Stemmer

Name of Database	Testing Time(Sec.)	Accuracy %	Precision %	Recall %	F-measure %
ACT-DS1	0.12	73%	74%	73%	73%
ACT-DS2	0.05	47%	51%	47%	48%
ACT-DS3	0.31	70%	70%	70%	70%
ACT-DS4	0.06	76%	77%	76%	75%
ACT-DS5	0.02	40%	41%	40%	39%
ACT-DS6	0.19	86%	86%	86%	86%
ACT-DS7	0.05	88%	88%	88%	88%
ACT-DS8	64.46	80%	80%	80%	80%
ACT-DS9	1632.23	91%	91%	91%	91%
ACT-DS10	0.06	85%	85%	85%	84%
ACT-DS11	0.05	48%	50%	48%	46%
Average	**154.33**	**71%**	**72%**	**71%**	**71%**

4.4.4. SVM Classifier Based on the Proposed Stemmer.

SVM proves in its results that it is a suitable classifier for Arabic document categorization, but in some datasets, it resulted in bad accuracy results. As you see, for example, in ACT-DS5 and ACT-DS2 the accuracy results were not satisfactory. Due to the nature of the documents. Because these methods do not take into account the classification of words that carry more than one meaning or synonyms. This is what prompted us to think differently to classify documents depending on the true sense of the word.

Table23
SVM with Our Stemmer

Name of Database	Testing Time	Accuracy	Precision	Recall	F-measure
ACT-DS1	0.67	93%	93%	93%	92%
ACT-DS2	0.34	76%	77%	76%	73%
ACT-DS3	1.2	91%	91%	91%	91%
ACT-DS4	0.15	88%	88%	88%	88%
ACT-DS5	0.07	56%	50%	56%	48%
ACT-DS6	0.37	94%	94%	94%	94%
ACT-DS7	0.08	98%	98%	98%	98%
ACT-DS8	0.11	93%	93%	93%	93%
ACT-DS9	0.36	98%	98%	98%	98%
ACT-DS10	0.12	96%	96%	96%	96%
ACT-DS11	0.22	77%	80%	77%	76%
Average	**0.37**	**87%**	**87%**	**87%**	**86%**

4.4.5. Naive Bayes classifier based on the Proposed Stemmer

(Table 24) shows the results of applying the Naive Bayes classifier on the datasets after they were processed using our stemmer. A glance to the results, we find that there is a clear improvement in the accuracy of the results compared to the previous stemmers, and is an expected result because our stemmer addressed all situations of Arabic word. As you see, Naive Bayes with our stemmer achieves more than 90% accuracy in most of the datasets.

Table 24
Bayesian with Our Stemmer

Name of Database	Testing Time	Accuracy	Precision	Recall	F-measure
ACT-DS1	7.51	82%	83%	82%	82%
ACT-DS2	3.03s	76%	81%	76%	73%
ACT-DS3	23.25	87%	88%	87%	87%
ACT-DS4	3.89	89%	90%	89%	89%
ACT-DS5	1.74	48%	38%	48%	37%
ACT-DS6	7.38	90%	91%	90%	90%
ACT-DS7	1.92	94%	95%	95%	95%
ACT-DS8	2.16	90%	91%	90%	91%
ACT-DS9	8.27	97%	97%	97%	97%
ACT-DS10	3.52	93%	94%	93%	94%
ACT-DS11	4.09	70%	74%	70%	68%
AVERAGE	**6.37**	**83%**	**84%**	**83%**	**82%**

4.5. Evaluation of Arabic Text Classification Based On SVD with Cosine Similarity

In this section we apply the SVD model with cosine distance depending on five representation models TF, TF-IDF, TF-IDF-CF, Bigram and Trigram. The results in tables 25, 26, 27, 28 and 29 prove that the best representation model with cosine distance is TF-IDF-CF. Because the TF-IDF-CF classified texts based on high frequency words in documents that belong to the same class, this method implies repetition represents document and category at once. Using trigram, as representation model, resulted in the worst results. This is sometimes the representation using (Trigram) were not reflected in the several documents that belong to the same category or may be contained in varying proportions, and this will weakens the ability to categorize it.

Table 25
Results of Applying the Cosine Similarity with TF

Name	**Cosine Similarity**			
of Database	**Accuracy**	**Recall**	**Precision**	**F-measure**
ATC-DS1	83%	82%	83%	82%
ATC-DS2	74%	69%	65%	67%
ATC-DS3	91%	91%	92%	91%
ATC-DS4	80%	81%	81%	81%
ATC-DS5	61%	57%	67%	62%
ATC-DS6	96%	95%	96%	95%
ATC-DS7	92%	92%	92%	92%
ATC-DS8	94%	93%	92%	92%
ATC-DS9	86%	85%	88%	86%
ATC-DS10	88%	89%	90%	89%
ATC-DS11	67%	74%	67%	70%
Average	**82%**	**82%**	**83%**	**82%**

Table 26
Results of Applying the Cosine Similarity with TF-IDF

Name of Database	Cosine similarity			
	Accuracy	**Recall**	**Precision**	**F-measure**
ATC-DS1	85%	83%	85%	84%
ATC-DS2	47%	53%	47%	50%
ATC-DS3	92%	92%	93%	92%
ATC-DS4	88%	88%	89%	88%
ATC-DS5	38%	44%	38%	41%
ATC-DS6	90%	87%	91%	89%
ATC-DS7	94%	94%	94%	94%
ATC-DS8	89%	89%	89%	89%
ATC-DS9	79%	77%	83%	80%
ATC-DS10	93%	94%	93%	93%
ATC-DS11	65%	66%	64%	65%
Average	**78%**	**79%**	**79%**	**79%**

Table 27
Results of Applying the Cosine Similarity with TFIDF-CF

Name of Database	Cosine similarity			
	Accuracy	**Recall**	**Precision**	**F-measure**
ATC-DS1	89%	87%	89%	88%
ATC-DS2	80%	7800%	82%	162%
ATC-DS3	96%	95%	96%	95%
ATC-DS4	88%	89%	89%	89%
ATC-DS5	84%	81%	87%	84%
ATC-DS6	97%	97%	97%	97%
ATC-DS7	100%	100%	100%	100%
ATC-DS8	96%	96%	95%	95%
ATC-DS9	92%	91%	92%	91%
ATC-DS10	95%	96%	95%	95%
ATC-DS11	80%	81%	86%	83%
Average	**91%**	**79%**	**92%**	**85%**

Table 28
Results of Applying the Cosine Similarity with Bigram

Name of Database	Cosine similarity Accuracy	Recall	Precision	F-measure
ATC-DS1	62%	54%	53%	53%
ATC-DS2	29%	43%	31%	36%
ATC-DS3	69%	69%	72%	70%
ATC-DS4	78%	76%	78%	77%
ATC-DS5	59%	58%	66%	62%
ATC-DS6	82%	80%	82%	81%
ATC-DS7	90%	90%	90%	90%
ATC-DS8	89%	85%	89%	87%
ATC-DS9	68%	65%	68%	66%
ATC-DS10	75%	76%	81%	78%
ATC-DS11	42%	48%	35%	40%
Average	**68%**	**68%**	**68%**	**68%**

Table 29
Results of Applying the Cosine Similarity with Trigram

Name of Database	Cosine similarity Accuracy	Recall	Precision	F-measure
ATC-DS1	67%	56%	55%	55%
ATC-DS2	53%	8%	44%	14%
ATC-DS3	71%	62%	63%	62%
ATC-DS4	78%	70%	77%	73%
ATC-DS5	66%	66%	66%	66%
ATC-DS6	83%	76%	86%	81%
ATC-DS7	87%	87%	86%	86%
ATC-DS8	75%	67%	80%	73%
ATC-DS9	68%	65%	69%	67%
ATC-DS10	64%	68%	65%	66%
ATC-DS11	23%	20%	28%	23%
Average	**67%**	**68%**	**65%**	**66%**

(Table 30) shows the results of applying the SVD on our dataset, note that Islam, health and sports classification accuracy rate achieved 100% percent, due to the text files in the corpus contains words with high repetition. Unlike politics and culture, here text files contain words that are used in more than one area which disperses some files and reduces the accuracy of categorization. For example, some words in literary texts may carry a different meaning to the word like poetic texts.

Table 30
Applying the SVD on our Dataset

Name of Category	Precision	Recall	F-measure
Art	93%	100%	96%
Sport	100%	100%	100%
Economy	90%	100%	94%
Law	100%	96%	98%
Health	100%	100%	100%
Technology	100%	92%	96%
Politic	93%	96%	95%
Islamic	100%	96%	98%
Education	95%	90%	93%
Culture	91%	91%	91%
Ideas	93%	90%	91%
Average	**96%**	**96%**	**96%**

4.6. Comparison between Similarity Measures Based on SVD

We examined all types of SVD with three similarity measure to find which combination gives the best results.

Table 31
Comparison between Similarity Measures

Similarity Measure	TF	TF-IDF	TF-IDF-CF	Bigram	Trigram
Euclidean Distance	57%	62%	80%	62%	**95%**
Manhattan Distance	59%	60%	78%	62%	**96%**
Hassanat's Distance	59%	60%	79%	61%	**94%**
Cosine Similarity	91%	92%	**96%**	69%	71%

From (Table 31), we can notice that the best accuracy can be achieved by using SVD of type TF-IDF-CF with cosine similarity measure. (Tables 25 to 30) show the detailed results for each dataset that include Recall, Precision, and F-measure for all SVD with cosine measure combinations.

Table 32
Results of Applying the Used Datasets over six Classifiers

Name of Dataset	(DT)	(NB)	(SVM)	(NN)	(KNN)	SVD with TF-IDF-CF
ACT-DS1	73%	82%	93%	65%	54%	89%
ACT-DS2	47%	76%	76%	41%	35%	80%
ACT-DS3	70%	87%	91%	41%	16%	96%
ACT-DS4	76%	89%	88%	53%	38%	88%
ACT-DS5	40%	48%	56%	39%	37%	84%
ACT-DS6	86%	90%	94%	83%	75%	97%
ACT-DS7	88%	94%	98%	82%	78%	100%
ACT-DS8	80%	90%	93%	76%	71%	96%
ACT-DS9	91%	97%	98%	85%	77%	92%
ACT-DS10	85%	93%	96%	61%	41%	95%
ACT-DS11	48%	70%	77%	46%	43%	80%
Average	**71%**	**83%**	**87%**	**61%**	**51%**	**91%**

(Table 32) summarizes the results of applying the used datasets over five classifiers, as well as, SVD. As can be seen from the table that our methodology gave better results than the other classifiers. If you look at ACT-DS7, you will find that SVD method achieved 100% accuracy rate, while the lowest ranked on the same data is KNN with accuracy percentage (78.2%).

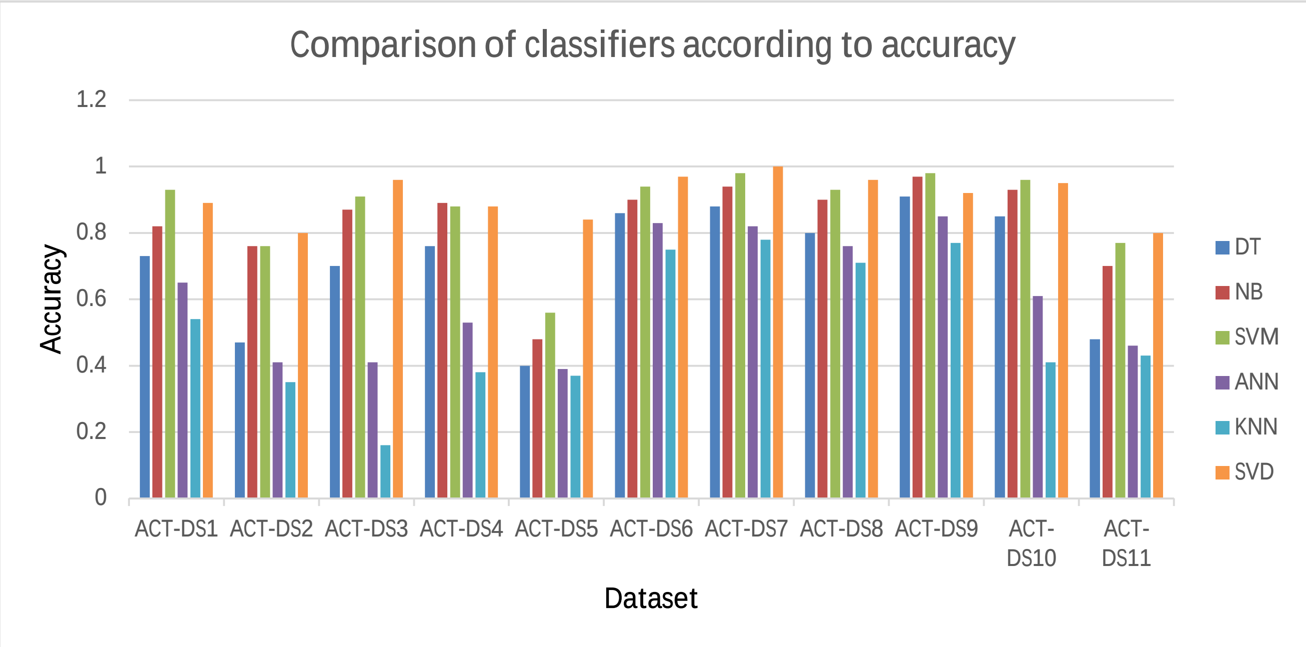


Figure 26
Comparison of Classifiers According to Accuracy

It is worthily mentioned here that the impact of the using of SVD in categorization was evident as in ACT-DS5, where classification accuracy rate was improved with a percentage ranged from 30-40%, this indicates that the SVD has dealt with texts in a quite differently way from other classifiers.

4.7Chapter six:Conclusions and Future Work

4.7.1 Conclusion

With increasing number of users of the Arabic language, as well as the diversity of communication channels which increased the opportunity to share ideas about current issues as a result of the evolution of the political situation in the countries, this led to the need to improve and increase performance efficiency of document categorization. Document categorization has become an urgent necessity. In this thesis, we have devised a method of categorization of Arabic texts to get high resolution compared to previous research.

We argue that preprocessing is an essential and urgent step before categorization to guarantee efficient results. Consequently, good preprocessing resulted in more meaningful and successful conclusions for end-users. Then, we construct four schemas that carry out efficient categorization. The findings of our experiments are summarized as follows: During our research, we did not find data set for experiments, as well as the available data sets is unbalanced, in this respect; we have developed our own corpora, namely ATC-DS3, a balanced dataset available online for research purposes.

There are differences between supervised classifiers in terms of categorization accuracy and consumed time. Through our experiments, we found that SVM and DT were the best in terms of accuracy, but DT needs a long time in execution. Results also showed that the NB and KNN are best in small documents. However, for large documents SVM is outperformed in document categorization. When applying F-measure on various data sets and different classifiers, SVM got the highest percentage of F-measure and Accuracy in all data sets. This is an expected result because large documents are commensurate with the nature of SVM. Unlike previous researches which mostly use precision and recall only, we try to test all classifiers using four performance measures; accuracy, F-measure, recall and precision. This criterion guarantees a comprehensive and detailed analysis of results.

In this work the concluded that removing the stop words led to better results in categorization by reducing the time consumed. At the beginning of the work we assumed that remove stop words will contribute to reduce the time consumed, but we noticed that the removing operation have a

positive impact on the categorization accuracy, for example when compare the results of Database Before and after removing of stop words

We noted that most researchers use root or light stemmer, but in this work we merged them determination hybrid stemmer integrates the advantages of each, and we got good results compared with previously.

The main contribution presented in this thesis focuses on using a hybrid method by merging SVD, TFIDF-CF and cosine distance. This hybrid method gave good results equivalent to SVM algorithm. We have shown that SVD techniques with n-gram and cosine distance presented the best performance in almost all data sets. The results of the experiments showed us that dimensionality reduction improved the categorization process in two ways: better results and better efficiency. Better results means that more documents are categorized correctly. Better efficiency means lower time complexity.

In addition, we perform an extensive set of experiments on three similarity measures; cosine distance, Euclidian distance and Hassanat's distance. It is noteworthy that Hassanat distance used for the first time in document categorization we conclude that the Hassanat distance gives better results than Euclidean, but cosine were the best.

We noticed that category had a role in the categorization accuracy, for example scientific documents give more accurate results over literary ones, this observation could be explained that scientific texts use clear words and fixed connotations, unlike literary texts that may carry completely different meaning to the word itself.

4.7.2 Future work

Arabic texts usually contain major and subcategory categorization, in our next development, we will improve our methodology to deal with multi-label documents. To classify texts accurately into main category and subcategory, A few researchers address this field, our high results on the one-label documents, gives us indicator that it will give good results also in the case of the multi-label. Also we will investigate hierarchical categorization of Arabic documents

In this study, we used the performance measures used by all researchers, but to be more precise, evaluation criteria must depend on the categorization algorithm. That's what we'll try to apply to our methodology. The future work will focus on studying impact of the type of language used in the accuracy of the results, so we'll try our methodology to other languages such as English and Turkish, also we will try more Arabic datasets.

We look forward to modify the proposed approach so as to overcome the problem of unbalanced data sets which is commonly encountered in

document categorization problems. And standardizing the lengths of documents and apply the same model.

Finally, Polysemy and synonymy are two problems encountered Arabic language. The first problem is that one word that can carry more than one meaning, and the second problem is that there is more than one word may carry the same meaning, and this would adversely affect the accuracy of categorization . This will require our attention in the context of a word within a sentence; we tried to use n-gram to minimize the impact of these problems. In the future we will consider the effective way to control these challenges.

References

Ababneh J., Almomani, O., Hadi, W., El-Omari, N. K. T., & Al-Ibrahim, A. 2014.Vector Space Models to Classify Arabic Text. s.l. : Ababneh, J., Almomani, O., Hadi, W., El-Omari, N. K. T., & Al-Ibrahim, A. (2014). Vector Space Models to Classify Arabic Text. International Journal of Computer Trends and Technology (IJCTT), 7(4), 219-223., 2014, **International Journal of Computer Trends and Technology (IJCTT),** 7(4), pp. 219-223.

Ababneh M, Al-Shalabi, R., Kanaan, G., & Al-Nobani, A. 2012. Building an Effective Rule-Based Light Stemmer for Arabic Language to Improve Search Effectiveness. **International Arab Journal of Information Technology (IAJIT)**, 9(4.

Ababneh, J, Almomani, O., Hadi, W., El-Omari, N. K. T., & Al-Ibrahim, A. 2014. Vector Space Models to Classify Arabic Text. **International Journal of Computer Trends and Technology (IJCTT),** 7(4, pp. 219-223.

Abbas M, & Berkani, D. 2006.Topic identification by statistical methods for arabic language. **WSEAS Transactions on Computers**, 5(9), pp. 1908-1913.

Abbas M., Smaili, K., & Berkani, D. 2011. Evaluation of Topic Identification Methods for Arabic Texts and their Combination by using a Corpus Extracted from the Omani Newspaper Alwatan. **Arab Gulf Journal of Scientific Research** 29.3-4, pp. 183-191.

Abbas, M, Smaili, K., & Berkani, D. 2010. Tr-classifier and knn evaluation for topic identification tasks. **The International Journal on Information and Communication Technologies (IJICT)**, 3(3), pp. 65-74.

Abbas, M., & Smaili, K. 2005. Comparison of topic identification methods for arabic language. In Proceedings of International Conference on Recent Advances in Natural Language Processing, **RANLP.** pp. 14-17.

Abbas, M., Smaili, K., & Berkani, D. 2009.Comparing TR-Classifier and KNN by using Reduced Sizes of Vocabularies. **Culture**, 1, 210.

Abbas, M., Smaili, K., & Berkani, D. 2009.Multi-category support vector machines for identifying Arabic topics. Advances in Computational Linguistics, **Special issue of Journal of Research in computing Science**, pp. 217-226.

Abu-Errub, A. 2014.Arabic Text Classification Algorithm using TFIDF and Chi Square Measurements. **International Journal of Computer Applications,** 93(6).

Addustour. [Online] http://www.addustour.com, .

Agarwal, S, Verma, A K and Dixit, N. 2014. Content Based Image Retrieval using Color Edge Detection and Discrete Wavelet Transform. s.l. : I**EEE, Issues and Challenges in Intelligent Computing Techniques (ICICT), International Conference** on. pp. 368-372.

Ahmad Al-falahi, Mohammed Al-sarem, Mohammed Ramdani, M. Bellafkih. 2014.Identify author of Arabic poem . **Conference: 5th International Conference on Arabic Language Processing**, At Oujda, Morocco, pp. 30-37.

Ahmed, N, Natarajan, T and Rao, K R. 1974. Discrete cosine transform. 1, **Computers, IEEE Transactions** on, Vol. 100, pp. 90-93.

Al gaad newspaper. [Online] http://www.alghad.com/.

AL Sulaiti, L. 2009. [Online] 2009. [Cited: November 15, 2015.] http://www.comp.leeds.ac.uk/eric/latifa/research.htm .

Alahali newspaper. [Online] http://www.hashd-ahali.net/main/ahali.

Alahmadi, A., Joorabchi, A., & Mahdi, A. E. 2013. **Combining Bag-of-Words and Bag-of-Concepts representations for Arabic text classification**. In Irish Signals & Systems Conference 2014 and 2014 China-Ireland International Conference on Information .

Al-Alwani, A., & Beseiso, M. 2013.Arabic Spam filtering using Bayesian Model. **International Journal of Computer Applications** 79(7), pp. 11-14.

Alanbat newspaper. [Online] http://www.alanbatnews.net, .

Alarabiawm . [Online] http://www.Alarabiawm.net.

Al-Arfaj, A., & Al-Salman,. 2015.Arabic NLP tools for ontology construction from Arabic text: An overview. In Electrical and Information Technologies (ICEIT), **International Conference on IEEE**. pp. 246-251.

Al-Gharaibeh, A., Al-Taani, A., & Alsmadi, I. 2011.The Usage of Formal Methods in Quran Search System. **In Proceedings of International Conference on Information and Communication Systems**. pp. 22-24.

Al-Harbi S, Almuhareb, A., Al-Thubaity, A., Khorsheed, M. S., & Al-Rajeh, A. 2008. **Automatic Arabic text classification**.

Aljlayl, M., & Frieder, O. 2002.On Arabic search: improving the retrieval effectiveness via a light stemming approach. In **Proceedings of the eleventh international conference on Information and knowledge management (ACM)**. pp. 340-347 .

Al-Kabi M. N, Wahsheh H. A., & Alsmadi I. M. 2014. **A Topical Classification of Hadith Arabic Text.** Amman, Jordan : s.n., ndInternational Conference on Islamic Applications in Computer Science And Technologyt.

Al-Kabi N. M, Kazakzeh, S. A., Ata, B. M. A., Al-Rababah, S. A., & Alsmadi, I. M. 2015. A novel root based Arabic stemmer. **Journal of King Saud University-Computer and Information Sciences**, 27(2, pp. 94-103.

Al-Kabi, M., Wahsheh, H., Alsmadi, I., Al-Shawakfa, E., Wahbeh, A., & Al-Hmoud, A. 2012.Content-based analysis to detect Arabic web spam. **Journal of Information Science** 38.3 , pp. 284-296.

Al-Kharashi, I. A., & Evens, M. W. 1994.Comparing words, stems, and roots as index terms in an Arabic information retrieval system. **Journal of the American Society for Information Science**, 45(8), pp. 548-560.

Almaany. [Online] 2015. [Cited: 11 15, 2015.] http://www.almaany.com/.

Al-Molegi A, Alsmadi, I., Najadat, H., Albashiri, H., Agrawal, K., & Singh, R. 2015. **Automatic Learning of Arabic Text Categorization.**

Al-Nashashibi, M. Y. A. 2014.Arabic Language Processing for Text Classification. Contributions to Arabic Root Extraction Techniques, Building An Arabic Corpus, and to Arabic Text Classification Techniques. s.l. : University of Bradford, 2014. pp. Al-Nashashibi, M. Y. A. (2014). **Arabic Language Processing for Text Classification. Contributions to Arabic Root Extraction Techniques, Building An Arabic Corpus, and to Arabic Text Classification Techniques** (Doctoral dissertation, University of Bradford).

Al-Omari, A., & Abuata, B. 2014.Arabic Light Stemmer (ARS). **Journal of Engineering Science and Technology**, 9(6), pp. 702-717.

Alqawasmeh, Iyas, Fawzi Harraq, Abd-almalek Salman. 2015.**Feature Extraction of Arabic text classification texts of Prophet Mohammed sayings**. s.l. : [26] استخراج الميزات لتصنيف نصوص الأحاديث النبوية الشريفة,, 2015, 12th ACS/IEEE International Conference on Computer Systems and Applications AICCSA .

Al-Radaideh, Q. A., & Al-Eroud, A. F. 2011. **Arabic Alert-Email Detection Using Rule Based Filter.**

Alrai newspaper. [Online] www.alrai.com.

Al-Ramahi, M. A., & Mustafa, S. H. 2012.**N-Gram-Based Techniques for Arabic Text Document Matching; Case Study: Courses Accreditation.**

Alsaleem, S. 2011. Automated Arabic Text Categorization Using SVM and NB. Int. **Arab J. e-Technol.**, 2(2, pp. 124-128.

Al-sarrayrih H, Al-Shalabi. R. 2009. **4th International Conference on Information**. pp. Clustering Arabic Documents Using Frequent Itemset-based Hierarchical Clustering with an N-Grams.

Al-Shammari, E., & Lin, J. 2008. **A novel Arabic lemmatization algorithm.** In Proceedings of the second workshop on Analytics for noisy unstructured text data. pp. 113-118.

Alshehri, A. M. 2002. Optimization and effectiveness of n-grams approach for indexing and retrieval in Arabic information retrieval systems.

Alsmearat, K., Al-Ayyoub, M., & Al-Shalabi, R. 2014. An extensive study of the Bag-of-Words approach for gender identification of Arabic articles. In Computer Systems and Applications (AICCSA), **IEEE/ACS 11th International Conference** . pp. 601-608.

Al-Taani A. T, & Al-Gharaibeh A. M. 2011. **Searching Concepts and Keywords in the Holy Quran.**

Al-Taani, A. T., & Al-Awad, N. A. K. 2009. An Empirical Analysis of Arabic WebPages Classification using Fuzzy Operators. **International Journal of Computational Intelligence** 5.1.

Al-Thubaity, A, Alhoshan, M., & Hazzaa, I. 2015. Using Word N-Grams as Features in Arabic Text Classification. s.l. : Springer International Publishing, In Software Engineering, Artificial Intelligence, **Networking and Parallel/Distributed Computing**, pp. 35-43.

Ameen, Abd-almalik. 2013. Approach to determine the language automatically (without supervision) in the city of multilabel texts. **Institutional Digital Repository for Naif Arab University for Security Sciences**, pp. 29-39.

Applied software design. [Online] 2012. [Cited: 11 15, 2015.] http://www.appliedsoftwaredesign.com/archives/cosine-similarity-calculator..

Arabic Corpora. [Online] 2015. https://sites.google.com/site/mouradabbas9/corpora .

Arabicstopwords. [Online] [Cited: November 15, 2015.] http://arabicstopwords.sourceforge.net.

Assabeel newspaper. [Online] http://www.assabeel.net.

Atwan, J., Mohd, M., Kanaan, G., & Bsoul, Q. 2014. Impact of Stemmer on Arabic Text Retrieval. s.l. : Atwan, J., Mohd, M., Kanaan, G., & Bsoul, Q. (2014). Impact of Stemmer on Arabic Text Retrieval. In Information Retrieval Technology (pp. 314-326). Springer International Publishing., 2014, In Information Retrieval Technology,. **Springer International Publishing**., pp. 314-326.

Awajan, A. 2007. Arabic text preprocessing for the natural language processing applications. **Arab Gulf Journal of Scientific Research**, pp. 179-189.

Awang Iskandar, D N, Thom, J A and Tahaghoghi, S M. 2008. Content-based image retrieval using image regions as query examples. **2Proceedings of the nineteenth conference on Australasian database.** Vol. 75, pp. 38-46.

BAEZA, Y., & Ribeiro-Neto, B. 2011.**Modern Information Retrieval-The Concepts and Technology behind Search.**

Basu, M. 2002. Gaussian-based edge-detection methods-a survey. 3, **IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS**, Vol. 32, pp. 252-260.

Bellet, A, Habrard, A and Sebban, M. 2013. A survey on metric learning for feature vectors and structured data. **arXiv preprint arXiv:**1306.6709. 2013.

Benala, T R, et al. 2014.Software Effort Estimation Using Data Mining Techniques. s.l. : Springer International Publishing, **ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India**-Vol I. pp. 85-92.

Bentley, P M and McDonnell, J T. 1994.Wavelet transforms: an introduction. **Electronics & communication engineering journal**, Vol. 6, pp. 175-186.

Bluebit. [Online] [Cited: November 15, 2015.] http://www.bluebit.gr/matrix-calculator.

Boll, S, Klas, W and Sheth, A P. 1998. Overview on Using Metadata to Manage Multimedia Data. **Multimedia Data Management: Using Metadata to Integrated and Apply Digital Data**. 1998, pp. 1-23.

Boomilingam, T and Subramaniam, M. 2014.Review on CBIR Trends and Techniques to Upgrade Image Retrieval. 7, **International Review on Computers and Software**, Vol. 9, pp. 1227-1240.

Brand, M. 2006.Fast low-rank modifications of the thin singular value decomposition. **Linear algebra and its applications**, 415(1), pp. 20-30.

Bruce, L M, Koger, C H and Li, J. 2002.Dimensionality reduction of hyperspectral data using discrete wavelet transform feature extraction. 10, **Geoscience and Remote Sensing, IEEE Transactions** on, Vol. 40, pp. 2331-2338.

Brunelli, R and Mich, O. 2001.Histograms analysis for image retrieval. 8, **Pattern Recognition**, Vol. 34, pp. 1625-1637.

Cambridge, AT&T Laboratories. 1992-1994.**ORL database for faces.** [Online] http://www.cl.cam.ac.uk/research/dtg/ attarchive/ facedatabase.html.

Canny, J. 1986. A computational approach to edge detection. 6, Pattern Analysis and Machine Intelligence, **IEEE Transactions** on, pp. 679-698.

Cha, S H. 2007.Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. 4, **INTERNATIONAL JOURNAL OF MATHEMATICAL MODELS AND METHODS IN APPLIED SCIENCES**, Vol. 1, pp. 300-307.

Chantar, H. K., & Corne, D. W. 2011. Feature subset selection for Arabic document categorization using BPSO-KNN. In **Nature and Biologically Inspired Computing (NaBIC), Third World Congress on IEEE,** pp. 546-551.

Chen, A., & Gey, F. C. 2002.Building an Arabic Stemmer for Information Retrieval. s.l. : **Building an Arabic Stemmer for Information Retrieval. In TREC** (Vol. 2002, pp. 631-639)., **In TREC,** pp. 631-639.

Cherif, W., Madani, A., & Kissi, M. 2014. Building a syntactic rules-based stemmer to improve search effectiveness for arabic language. In Intelligent Systems: Theories and Applications (SITA-14), **9th International Conference on IEEE**. pp. 1-6.

Cherif, W., Madani, A., & Kissi, M. 2014. Integrating effective rules to improve arabic text stemming. In Multimedia Computing and Systems (ICMCS), **International Conference on IEEE**. pp. 1077-1081.

Choras, R S. 2007.Feature extraction for CBIR and biometrics applications. **7th WSEAS International Conference on Applied Computer Science.** Vol. 7.

Chowdhury, M, Das, S and Kundu, M K. 2013. A ripplet transform based statistical framework for natural color image retrieval. **Image Analysis and Processing–ICIAP**, pp. 492-502.

Clough, P, et al. 2005. **The CLEF** cross-language image retrieval track. Vienna : s.n., CLEF Workshop.

Dan, L., Lihua, L., & Zhaoxin, Z. 2013. Research of text categorization on WEKA. In Intelligent System Design and Engineering Applications (ISDEA), **Third International Conference on IEEE**. pp. 1129-1131.

Darwish, K., & Oard, D. W. 2003. CLIR Experiments at Maryland for TREC-2002: Evidence combination for Arabic-English retrieval. (No. LAMP-TR-101). **MARYLAND UNIV COLLEGE PARK INST FOR ADVANCED COMPUTER STUDIES**.

Davis, L S. 1975.A survey of edge detection techniques. **Computer graphics and image processing,** Vol. 4, pp. 248-270.

Deb, S and Zhang, Y. 2004.An overview of content-based image retrieval techniques. s.l. : IEEE, Advanced Information Networking and Applications, **AINA 18th International Conference** on. pp. 59-64.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. 1990. **Indexing by latent semantic analysis. JAsIs**, 41(6), pp. 391-407.

Deselaers, T, Keysers, D and Ney, H. 2004. Features for image retrieval: A quantitative comparison. **Springer Berlin Heidelberg : s.n., Pattern Recognition,** pp. 228-236.

Deselaers, T, Keysers, D and Ney, H. 2008.Features for image retrieval: an experimental comparison. 2, **Information Retrieval,** Vol. 11, pp. 77-107.

Deshpande, G and Borse, M. 2011. Image Retrieval with the use of different color spaces and the texture feature. Int. **Conf. Soft. Comput. Appl**, pp. 273-278.

Drucker, H., Wu, D., & Vapnik, V. N. 1999. Support vector machines for spam categorization. Neural Networks,**IEEE Transactions** on, 10(5), pp. 1048-1054.

Dunning, T. 1994.**Statistical identification of language.** s.l. : Computing Research Laboratory, New Mexico State University., 1994.

Duwairi R, Al-Refai, M., & Khasawneh, N. 2007. Stemming versus light stemming as feature selection techniques for arabic text categorization. **In Innovations in Information Technology**, **IIT'07. 4th International Conference**. pp. 446-45.

Duwairi, R. M. 2015. Sentiment analysis for dialectical Arabic. **Information and Communication Systems (ICICS),** 6th International Conference IEEE, pp. 166-170.

ElAlami, M E. A new matching strategy for content based image retrieval system. **Applied Soft Computing**, Vol. 14, pp. 407-418.

Elayidom, M. S., Jose, C., Puthussery, A., & Sasi, N. K. 2013. Text classification for authorship attribution analysis. **arXiv preprint arXiv:**1310.4909.

Elbery, A. 2012.**Classification of Arabic Documents.**

El-Halees, A. 2009. Filtering spam e-mail from mixed arabic and english messages: a comparison of machine learning techniques. **Int. Arab J. Inf. Technol**, pp. 52-59.

Elhassan, R., & Ahmed, M. 2015. Arabic Text Classification on Full Word. **International Journal of Computer Science and Software Engineering (IJCSSE),** pp. 114-120.

El-Khair, I. A. 2006. Effects of stop words elimination for Arabic information retrieval: a comparative study. **International Journal of Computing & Information Sciences**, 4(3), pp. 119-133.

ElSahar, H., & El-Beltagy, S. R. 2015.Building Large Arabic Multi-domain Resources for Sentiment Analysis. **Computational Linguistics and Intelligent Text Processing Springer International Publishing**, pp. 23-34.

Fogel, I and Sagi, D. 1989.Gabor filters as texture discriminator. 2, **Biological cybernetics**, Vol. 61, pp. 103-113.

Ford, A and Roberts, A. 1998. **Colour space conversions.** London : s.n., Westminster University, pp. 1-31.

Foster, D. W. 2000. **Author unknown: On the trail of anonymous**. Macmillan.

Free dictionary. [Online] 2015. [Cited: 11 15, 2015.] http://ar.thefreedictionary.com/.

Fu, K S and Mui, J K. 1981. A survey on image segmentation. **Pattern recognition,** Vol. 13, pp. 3-16.

Gabor, D. Theory of communication. Part 1: The analysis of information. 26, **Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering**, Vol. 93, pp. 429-441.

Gharib, T. F, Habib, M. B., & Fayed, Z. T. 2009. Arabic Text Classification Using Support Vector Machines. **IJ Comput. Appl**., 16(4), pp. 192-199.

Ghazi R, Riyad S ,Ghassan K,Asmaa N. 2015.Different Classification Algorithms Based on Arabic Text Classification: Feature Selection Comparative Study. **International Journal of Advanced Computer Science and Applications (IJACSA).**

Ghwanmeh, S., Kanaan, G., Al-Shalabi, R., Rabab'ah. 2009. Enhanced Algorithm for Extracting the Root of Arabic Words. In: the Sixth International Conference on Computer Graphics, Imaging and Visualization, **(CGIV '09)**. pp. 388–391.

Giannakopoulos, G., Mavridi, P., Paliouras, G., Papadakis, G., & Tserpes, K. 2012.**Representation Models for Text Classification: a comparative analysis over three Web document types.** In Proceedings of the 2nd International Conference on Web Inte.

Giveki, D, et al. 2015.A New Content Based Image Retrieval Model Based on Wavelet Transform. **Journal of Computer and Communications,** Vol. 3, pp. 66-73.

Gong, Y, Chuan, C H and Xiaoyi, G. 1996.Image indexing and retrieval based on color histograms. 2, **Multimedia Tools and Applications,** Vol. 2, pp. 133-156.

Gonzalez, R C and Woods, R E. 2002.**Digital image processing.**

Guo, Gongde, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. 2003. **KNN model-based approach in classification.** In On The Move to Meaningful Internet Systems: CoopIS, DOA, and ODBASE,, pp. pp. 986-996. Springer Berlin Heidelberg, .

Guyon, I and Elisseeff, A. 2006. An Introduction to Feature Extraction. **Feature Extraction.** pp. 1-25.

Haddi, E., Liu, X., & Shi, Y. 2013.The role of text pre-processing in sentiment analysis. **Procedia Computer Science** 17, pp. 26-32.

Hadni M, Lachkar, A., & Ouatik, S. A. 2012. A new and efficient stemming technique for Arabic Text Categorization. In Multimedia Computing and Systems (ICMCS), **International Conference on IEEE**. pp. 791-796.

Hadni, M., Lachkar, A., & Ouatik, S. A. A. A **Comparative Study with Different Feature Selection for Arabic Text Categorization**. pp. 96-101.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. 2009. The WEKA data mining software: an update. **ACM SIGKDD explorations newsletter**, 11(1), pp. 10-18.

Hamouda, K. Chantar, David W. Corne. 2011.Arabic text categorization via binary particle swarm optimization and support vector machines. Slovenia : s.n., in proceeding of the 5th Int'l Conf. **on Bio-Inspired Optimization Methods and their Application**s. pp. 301-310.

Hampel, F R, et al. 2011.Robust statistics: the approach based on influence functions. s.l. : **John Wiley & Sons,** Vol. 114.

Harrag F, El-Qawasmah E , & Al-Salman A. M. S. 2011.Stemming as a feature reduction technique for Arabic Text Categorization. April 128-133, **Programming and Systems (ISPS), 10th International Symposium on. IEEE**.

Harrag, F., & Al-Qawasmah, E. 2010.Improving Arabic Text Categorization Using Neural Network with SVD.. **JDIM**, 8(4), 233-239. : s.n., JDIM, 8(4), pp. 233-239.

Hassanat, A B. 2014.Dimensionality Invariant Similarity Measure. 8, **Journal of American Science**, Vol. 10, pp. 221-226.

Hassanat, A. B. 2014. 2014, **arXiv preprint arXiv:**1409.0923, pp. 221-226.

Hassanat, A. B. 2014. Dimensionality Invariant Similarity Measure. **arXiv preprint arXiv**:1409.0923.

Hiremath, P S, Shivashankar, S and Pujari, J. 2006.Wavelet based features for color texture classification with application to CBIR. **International Journal of Computer Science and Network Security**, Vol. 6, pp. 124-133.

Huang, A. 2008. Similarity measures for text document clustering. Christchurch, New Zealand : s.n., 2008. **In Proceedings of the sixth new zealand computer science research student conference** (NZCSRSC2008). pp. 49-56.

Huang, H. Y., Shih, W. S., & Hsu, W. H. 2007. Movie classification using visual effect features. **In Signal Processing Systems, IEEE Workshop on IEEE**. pp. 295-300.

Huang, J, et al. 1997.**Image indexing using color correlograms**. s.l. : IEEE, Computer Vision and Pattern Recognition Computer Vision and Pattern Recognition, Proceedings., IEEE Computer Society Conference on. pp. 762-768.

Huang, Z C, et al. 2010. Content-based image retrieval using color moment and Gabor texture feature. s.l. : **IEEE, Machine Learning and**

Cybernetics (ICMLC), 2010 International Conference on. Vol. 2, pp. 719-724.

Hussein, A. S. 2015.Arabic document similarity analysis using n-grams and singular value decomposition. In Research Challenges in Information Science (RCIS), **IEEE 9th International Conference on IEEE**. pp. 445-455.

Ibrahim, H. S., Abdou, S. M., & Gheith, M20. 2015. Sentiment Analysis For Modern Standard Arabic And Colloquial. **arXiv preprint arXiv**:1505.03105.

Jasmine, K P and Kumar, P R. 2014.Color Histogram and DBC Co-Occurrence Matrix for Content Based Image Retrieval. 6, **International Journal of Information Engineering and Electronic Business (IJIEEB)**, Vol. 6, p. 47.

Jhanwar, N, et al. 2004. Content based image retrieval using motif cooccurrence matrix. 14, **Image and Vision Computing**, Vol. 22, pp. 1211-1220.

Junod, J. 1997. Servers to spam: drop dead. Computers & Security. Khamar, Khushbu. 2013. Short Text Classification Using kNN Based on Distance Function. **IJARCCE International Journal of Advanced Research in Computer and Communication Engineering. Government Engineering College,** Modasa, pp. SSN Print: 2319-5940 ISSN Online: 2278-1021.

Kanan, T., & Fox, E. A. 2015.**Automated Arabic Text Classification with P-Stemmer, Machine Learning, and a Tailored News Article Taxonomy**. Department of Computer Science, Virginia Polytechnic Institute & State University.

Katakis, Ioannis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2010 .Tracking recurring contexts using ensemble classifiers: An application to email filtering. **Knowledge and Information Systems** 22, pp. 371-391.

Khalifa, Hamouda. **Arabic Datasets**. [Online] [Cited: November 15, 2015.] http://www.macs.hw.ac.uk/~hhc6/.

Khorsheed, M.S. and Al-Thubaity, A.O. 2013. Comparative evaluation of text classification techniques using a large diverse Arabic dataset. **Springer Science + Business Media Dordrecht.**

Kiritchenko, S. 2005.**Hierarchical text categorization and its application to bioinformatics** . s.l. : (Doctoral dissertation, University of Ottawa).

Kobayashi, K and Chen, Q. 2015.Content-Based Image Retrieval Using Features in Spatial and Frequency Domains. I**ntelligence in the Era of Big Data,** pp. 269-277.

Kodratoff, Y. 2014.**Introduction to machine learning.** s.l. : Morgan Kaufmann, 2014.

Krause, E. F. 1986.**Taxicab Geometry: An Adventure in Non-Euclidean Geometry.** New York: Dover.

Kuncheva, L. I., Whitaker, C. J., Shipp, C. A., & Duin, R. P. 2003.Limits on the majority vote accuracy in classifier fusion. **Pattern Analysis & Applications**, 6(1), pp. 22-31.

Lam, W., Ruiz, M., & Srinivasan, P. 1999.Automatic text categorization and its application to text retrieval. **Knowledge and Data Engineering, IEEE Transactions on**, 11(6). pp. 865-879.

Lamard, M, et al. 2007.Content based image retrieval based on wavelet transform coefficients distribution. s.l. : **IEEE, Engineering in Medicine and Biology Society, EMBS 29th Annual International Conference of the IEEE**. pp. 4532-4535.

Lancaster. [Online] 2015. [Cited: November 15, 2015.] http://www.lancaster.ac.uk/staff/elhaj/corpora.htm.

Larkey, L. S., Ballesteros, L., & Connell, M. E. 2002.**Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis**. In Proceedings of the 25th annual international ACM SIGIR conference on Research and developme.

Larkey, L. S., Ballesteros, L., & Connell, M. E. 2007. Light stemming for Arabic information retrieval. In **Arabic computational morphology, Springer Netherlands**, pp. 221-243.

Leauhatong, T, Atsuta, K and Kondo, S. 2007. A new content-based image retrieval using color correlogram and inner product metric. **Image Analysis for Multimedia Interactive Services**, 2007. WIAMIS'07. Eighth International Workshop on, pp. 33-33.

Lehmann, TM, et al. 2003. The IRMA code for unique classification of medical images. **SPIE.** Vol. 5033, pp. 440-451.

Lehmann, TM, et al. 2005.**irma-10000 database.** [Online] 1, 2005. https://ganymed.imib.rwth-aachen.de/irma/datasets_en.php?SELECTED=00004#00004.dataset.

Lei, Z, Fuzong, L and Bo, Z. 1999.A CBIR method based on color-spatial feature. s.l. : IEEE, TENCON 99. **Proceedings of the IEEE Region 10 Conference**. Vol. 1, pp. 166-169.

Lew, M S, et al. 2006. Content-based multimedia information retrieval: State of the art and challenges. **ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)**. Vol. 2, 1, pp. 1-19.

Li, C. H., & Park, S. C. 2007. Neural network for text classification based on singular value decomposition. In Computer and Information Technology. CIT 2007. **7th IEEE International Conference on IEEE**. pp. 47-52.

Lin, P Y. 2007. **An introduction to wavelet transform. Taipei Taiwan** : s.n., Graduate Institute of Communication Engineering National Taiwan University.

Liu, G H. 2015.Content-Based Image Retrieval Based On Visual Attention And The Conditional Probability. s.l. : Atlantis Press, **International Conference on Chemical, Material and Food Engineering (CMFE).** pp. 838-842.

Lumb, M and Sethi, P. 2013. Texture Feature Extraction of RGB, HSV, YIQ and Dithered Images using Wavelet and DCT Decomposition Techniques. 10, **International Journal of Computer Applications,** Vol. 73, pp. 41-49.

Lumb, M and Sethi, P. Texture Feature Extraction of RGB, HSV, YIQ and Dithered Images using GLCM, Wavelet Decomposition Techniques. 25-31. 11, 25-31, **International Journal of Computer Applications,** Vol. 68, pp. 25-31.

M. Abbas, K. Smaili, D. Berkani. 2011. Evaluation of Topic Identification Methods on Arabic Corpora. **Journal of Digital Information Management** Vol. 9 No. 5,, pp. 185-192.

Malmasi, S., Refaee, E., & Dras, M. 2015. Arabic Dialect Identification using a Parallel Multidialectal Corpus. s.l. : In **Proceedings of PACLING**, pp. 209-217.

Mandal, J. K., Suresh Chandra Satapathy, Manas Kumar Sanyal, Partha Pratim Sarkar, and Anirban Mukhopadhyay, eds. 2015. Information Systems Design and Intelligent Applications. **Proceedings of Second International Conference INDIA**. Vol. 1. Springer, 2015.

Manjunath, B S and Ma, W Y. 1996. Texture features for browsing and retrieval of image data. 8, s.l. : IEEE, Pattern Analysis and Machine Intelligence, **IEEE Transactions on**, Vol. 18, pp. 837-842.

Manjunath, B S, et al. 2001. Color and texture descriptors. 6, s.l. : IEEE, Circuits and Systems for Video Technology, **IEEE Transactions** on, Vol. 11, pp. 703-715.

Mebane, W R and Sekhon, J S. 2004. Robust estimation and outlier detection for overdispersed multinomial models of count data. 2, **American Journal of Political Science**, Vol. 48, pp. 392-411.

Miloslavsky, Alec, Jason Goecke, Vladimir N. Deryugin, Dmitry A. Torba, Igor Neyman, and Oleg Turovsky. 2008.**Routing emails to a plurality of agents.** 7460496 US , 2008.

Mingyong Liu, Jiangang Yang. 2012 .An improvement of TFIDF weighting in text categorization . **International Conference on Computer Technology and Science** (ICCTS 2012), pp. 44-47.

Mishra, P and Prajapati, G S. 2014.CONTENT BASED IMAGE RETRIEVAL THROUGH DISTANCE SIMILARITY METRICS. 2,

International Journal of Computer Science and Information Security, Vol. 12.
Moh’d Mesleh, A. 2008. Support vector machines based Arabic language text classification system: feature selection comparative study. s.l. : Springer Netherlands, In **Advances in Computer and Information Sciences and Engineering** , pp. 11-16.
Mohammad, I. A. 2010. Classical Arabic Poetry Categorization Using N-Gram Frequency Statistics. **Iraqi Journal of Science**51(1), pp. 159-165.
Mohanaiah, P, Sathyanarayana, P and GuruKumar, L. 2013. Image texture feature extraction using GLCM approach. 5, **International Journal of Scientific and Research Publications**, Vol. 3.
Moh'd A Mesleh, A. 2007.Chi square feature extraction based SVMs Arabic language text categorization system. **Journal of Computer Science**, 3(6), pp. 430-435.
Mountassir, A., Benbrahim, H., & Berrada, I. 2012.**Sentiment Classification on Arabic corpora: preliminary results of a cross-study.** 3 e Séminaire de Veille Stratégique, Scientifique et Technologique (VSST’12).
Müller, H, et al. 2001.Performance evaluation in content-based image retrieval: overview and proposals. 5, **Pattern Recognition Letters,** Vol. 22, pp. 593-601.
Murty, M. R, Murthy, J. V. R., & PVGD, P. R. (2011).Text Document Classification based-on Least Square Support Vector Machines with Singular Value Decomposition. **International Journal of Computer Applications**, 27(7).
Mustafa, S. H. 2012.Word Stemming for Arabic Information Retrieval: The Case for Simple Light Stemming. **Abhath Al-Yarmouk: Science & Engineering Series,** 21(1).
Nene, S A, Nayar, S K and Murase, H. 1996.**Coil-100 database.** [Online] 1996. http://www1.cs.columbia.edu/CAVE/software/softlib/coil-100.php.
Niranjanan, S and Gopalan, S R. 2012. Performance efficiency of quantization using HSV colour space and intersection distance in CBIR. 21, **International Journal of Computer Applications,** Vol. 42, pp. 48-55.
Odeh A, Abu-Errub, A., Shambour, Q., & Turab, N. 2015.Arabic Text Categorization Algorithm using Vector Evaluation Method. **arXiv preprint arXiv**:1501.01318.
Orchard, M T and Bouman, C. 1991.Color quantization of images. 12, s.l. : **IEEE, Signal Processing, IEEE Transaction**s on, Vol. 39, pp. 2677-2690.

Otair, M. 2013. Comparative analysis of Arabic stemming algorithms. **International Journal of Managing Information Technology (IJMIT)** Vol 5 , pp. 1-12.

Patel, O, Maravi, Y P and Sharma, S. 2013. A comparative study of histogram equalization based image enhancement techniques for brightness preservation and contrast enhancement. **arXiv preprint arXiv**:1311.4033.

Patel, R and Parmar, S. 2014. Content based image retrieval for computed tomography images using Support Vector Machine classifier. s.l. : **IEEE, Emerging Technology Trends in Electronics, Communication and Networking (ET2ECN), 2nd International Conference** on. pp. 1-6.

ping Tian, D. 2013. A review on image feature extraction and representation techniques. **International Journal of Multimedia and Ubiquitous Engineering**, Vol. 8, pp. 385-396.

Puviarasan, N and Bhavani, R. 2014.Retrieval of Images Using Weighted Features. 1, **International Journal of Advanced Computer Research,** Vol. 4, p. 60.

Qu, Y, et al. 2003. Data reduction using a discrete wavelet transform in discriminant analysis of very high dimensionality data. 1, **Biometrics**, Vol. 59, pp. 143-151.

Quinlan, J. Ross. 2014.programs for machine learning. s.l. : **Elsevier.**

Rao, A, Srihari, R K and Zhang, Z. 1999. Spatial color histograms for content-based image retrieval. s.l. : IEEE, Tools with Artificial Intelligence, Proceedings. **11th IEEE International Conference on.** pp. 183-186.

Rasli, R M, et al. 2012. Comparative Analysis of Content Based Image Retrieval Techniques Using Color Histogram: A Case Study of GLCM and K-Means Clustering. s.l. : **IEEE, Intelligent Systems, Modelling and Simulation (ISMS), Third International Conference** on. pp. 283-286.

Ravishankar, K C, et al. 1999.Dominant color region based indexing for cbir. s.l. : **IEEE, Image Analysis and Processing, Proceedings. International Conference** on. pp. 887-892.

Rosenblatt, Frank. 1962.**Principles of neurodynamics; perceptrons and the theory of brain mechanisms**. Washington : Spartan Books.

Rotating screw. [Online] [Cited: July 15, 2015.] http://www.rotatingscrew.com/utfcast-professional.aspx.

Ruiz M. E., & Srinivasan, P. 1998.**Automatic text categorization using neural networks.** In Proceedings of the 8th ASIS SIG/CR Workshop on Classification Research, pp. 59-72.

Ruiz, M. E., & Srinivasan, P. 1988. **Automatic text categorization using neural networks**. In Proceedings of the 8th ASIS SIG/CR Workshop on Classification Research. pp. 59-72.

S. Khoja, R. Garside. 1999. Stemming Arabic Text. [Online] 1999. [Cited: 11 15, 2015.] http://www.comp.lancs.ac.uk/computing/users/khoja/stemmer.

Saad, M. K. 2010.The impact of text preprocessing and term weighting on Arabic text classification. Gaza : The Islamic University, 2010. pp. Saad, M. K. (2010). **The impact of text preprocessing and term weighting on Arabic text classification** (Doctoral dissertation, The Islamic University-Gaza).

Saad, M. K., & Ashour, W. 2010. OSAC: Open Source Arabic Corpora. Cyprus : s.n., 2010. In 6th **International Symposium on Electrical and Electronics Engineering and Computer Science.** pp. 118-123.

Saha, S K, Das, A K and Chanda, B. 2004. Cbir using perception based texture and colour measures. s.l. : **IEEE, Pattern Recognition, ICPR Proceedings of the 17th International Conference** on. Vol. 2, pp. 985-988.

Said D, Wanas, N. M., Darwish, N. M., & Hegazy, N. 2009.A study of text preprocessing tools for arabic text categorization. **In The second international conference on Arabic language**. pp. 230-236.

Sawalha, M., & Atwell, E. S. 2008.Comparative evaluation of arabic language morphological analysers and stemmers. In Proceedings of COLING, 22nd I**nternational Conference on Comptational Linguistics (Poster Volume) Coling Organization**. pp. 107-110.

Shambharkar, P.G. and Doja, M.N. 2015.**Automatic categorization of movie trailers using data mining techniques: A review.** Computing, Communication & Automation (ICCCA). pp. 88-94.

Sharma, M and Batra, A. 2014. Analysis of distance measures in content based image retrieval. **global journal of computer science and technology**, Vol. 14.

Sia, K. C., Cho, J., & Cho, H. K. 2007. Efficient monitoring algorithm for fast news alerts. **Knowledge and Data Engineering, IEEE Transactions** on, 19(7), pp. 950-961.

Singh, S M and Hemachandran, K. 2012. Image Retrieval based on the combination of Color Histogram and Color Moment. 3, **International Journal of Computer Applications**, Vol. 58.

Skiena, S. 1990. Implementing Discrete Mathematics: Combinatorics and Graph Theory with Mathematica. **Reading, MA: Addison-Wesley**, p. 172 and 227.

Smeulders, A W, et al. 2000. Content-based image retrieval at the end of the early years. 12, **Pattern Analysis and Machine Intelligence, IEEE Transactions** on, Vol. 22, pp. 1349-1380.

Smith, C. 2013. Facebook Users Are Uploading 350 Million New Photos Each Day. **Business Insider**. [Online] 2013. [Cited: 6 12, 2015.] http://www.businessinsider.com/facebook-350-million-photos-each-day-2013-9.

Song, X, Li, Y and Chen, W. 2008.A Textural Feature-Based Image Retrieval Algorithm. s.l. : IEEE, **Natural Computation, ICNC'08. Fourth International Conference on**. Vol. 4, pp. 71-75.

Srivastava, P, Prakash, O and Khare, A. 2014. Content-Based Image Retrieval using moments of wavelet transform. s.l. : I**EEE, Control, Automation and Information Sciences (ICCAIS), International Conference** on. pp. 159-164.

Stop-Word. [Online] 2011. [Cited: 11 15, 2015.] https://code.google.com/p/stop-words/downloads/list.

Suhasini, P S, Krishna, K and Krishna, I M. 2009.CBIR USING COLOR HISTOGRAM PROCESSING. 1, **Journal of Theoretical & Applied Information Technology,** Vol. 6.

Syiam, Mostafa M., Zaki T. Fayed, and Mena B. Habib. 2006.An intelligent system for Arabic text categorization. **International Journal of Intelligent Computing and Information Sciences** 6.1, pp. 1-19.

Symeonidis, P., Kehayov, I., & Manolopoulos, Y. 2012.Text classification by aggregation of SVD eigenvectors. In Advances in Databases and Information Systems. January, **Springer Berlin Heidelberg,** pp. 385-398.

Taghva, K., Elkhoury, R., & Coombs, J. 2005. Arabic stemming without a root dictionary. In null . **IEEE**., pp. 152-157.

Tsutsumi, Kimitaka, Kazutaka Shimada, and Tsutomu Endo. 2007. Movie review categorization based on a multiple classifier. the 21th Pacific Asia Conference on Language, **Information and Computation (PACLIC).**

Uysal, A. K., & Gunal, S. 2014. The impact of preprocessing on text classification. **Information Processing & Management**, 50(1), pp. 104-112.

Vadivel, A K, Majumdar, A K and Sural, S. 2003. Performance comparison of distance metrics in content-based image retrieval applications. Bhubaneswar, India : s.n., **International Conference on Information Technology (CIT)**, pp. 159-164.

Van de Geer, J P. 1995.**Some Aspects of Minkowski Distances**. Leiden : Leiden University.

VirginiaTech. [Online] 2012. [Cited: 11 14, 2015.] http://vtechworks.lib.vt.edu/handle/10919/19096?show=full.

w3techs. **Features extraction to classify the texts of speech prophet muhammad** [Online] 25/6/2015. [Cited: 11 13, 2015.]

http://www.w3techs.com/technologies/overview/content_language/all.

Wahsheh, H. A., Al-Kabi, M. N., & Alsmadi, I. M. 2012.**Spam detection methods for Arabic web pages**. In First Taibah University International Conference on Computing and Information Technology-Information Systems ICCIT. pp. 486-490.

Walia, E, et al. 2014.Variant of completed Robust LBP for two-level probabilistic content based Image retrieval. **Proceedings of Computer and Advanced Technology in Education** http://dx. doi. org/10.2316/P, Vol. 3.

Wang, J Z.**Wang database**. [Online] http://wang.ist.psu.edu/docs/related/.

Wang, S. 2001.**A robust CBIR approach using local color histograms**. s.l. : University of Alberta.

Wang, Y. X., & Deng, W. W. 2009. A New Text Categorization Method Based on SVD and Cascade Correlation Algorithm. In Artificial Intelligence and Computational Intelligence, **AICI'09. International Conference on IEEE**. Vol. 3, pp. 57-60.

wikipedia. [Online] 8 22, 2015. [Cited: 11 13, 2015.] http://goo.gl/izzGEA.

Wu, P, et al. 1999. A texture descriptor for image retrieval and browsing. s.l. : **IEEE, Content-Based Access of Image and Video Libraries**, pp. 3-7.

Wu, P, et al. 2004.A texture descriptor for browsing and similarity retrieval. **Signal Processing: Image Communication**, Vol. 16, pp. 33-43.

Yadav, O and Suryawanshi, V. 2014.CBIR Evaluation using Different Distances and DWT. 16, **International Journal of Computer Applications,** Vol. 93.

Yang, M, Kpalma, K and Ronsin, J. 2008. **A survey of shape feature extraction techniques.** Pattern recognition, pp. 43-90.

Yousif, S. A., Samawi, V. W., Elkabani, I. S. L. A. M., & Zantout, R. A. C. H. E. D.Enhancement Of Arabic Text Classification Using Semantic Relations With Part Of Speech Tagger. **W transactions Advances In Electrical And Computer Engineering**, pp. 195-201.

Yu, H, et al. 2002. Color texture moments for content-based image retrieval. s.l. : **IEEE, Image Processing. Proceedings. International Conferenc**e on. Vol. 3, pp. 929-932.

Yu, L. C., Hu, H. L., & Lin, W. H. 2013.Evaluation on text categorization for mathematics application questions. **Asia-Pacific : s.n., In Signal and Information Processing Association Annual Summit and Conference (APSIPA)**. pp. 1-2.

Zaghoul, F. A, & Al-Dhaheri, S. 2013. Arabic Text Classification Based on Features Reduction Using Artificial Neural Networks. In Computer

Modelling and Simulation (UKSim), **UKSim 15th International Conference on IEEE**. pp. 485-490.

Zhang, D and Lu, G. 2001.Content-based shape retrieval using different shape descriptors: A comparative study. s.l. : **IEEE, ICME International Conference on Multimedia and Expo,IEEE International Conference** on . p. 289.

Zhang, D and Lu, G. 2002. A comparative study of three region shape. **Digital Image Computing Techniques and Applications DICTA,** pp. 1-6.

Zhang, D and Lu, G. 2003. Evaluation of similarity measurement for image retrieval. s.l. : **IEEE, Neural Networks and Signal Processing, Proceedings of the International Conference** on. Vol. 2, pp. 928-931.

Zhang, J, Li, G L and He, S W. 2008.Texture-based image retrieval by edge detection matching GLCM. s.l. : IEEE, High Performance Computing and Communications, **HPCC'08. 10th IEEE International Conference on**. pp. 782-786.

Zinner, T., Lemmerich, F., Schwarzmann, S., Hirth, M., Karg, P., & Hotho, A. 2015.Text Categorization for Deriving the Application Quality in Enterprises Using Ticketing Systems. **In Big Data Analytics and Knowledge Discovery, Springer Int,** pp. 325-336.

Zou, X and You, F C. 2015. An Image Retrieval Method Based on PGH and Distance Histogram. s.l. : Atlantis Press, **International Conference on Artificial Intelligence and Industrial Engineering**.

المعلومات الشخصية

الاسم: محمد محمود العوضي

التخصص: ماجستير علم الحاسوب

الكلية: العلوم

السنة: 2015

هاتف رقم: 0772366905